

REPORT DOCUMENTATION PAGE			Form Approved OMB No. 0704-0188	
Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of Management and Budget, Paperwork Reduction Project (0704-0188), Washington, DC 20503.				
1. AGENCY USE ONLY (Leave blank)		2. REPORT DATE 31.Jan.03		3. REPORT TYPE AND DATES COVERED DISSERTATION
4. TITLE AND SUBTITLE EXACT TEST SIZE AND POWER FOR SMALL SAMPLES USING AN INTERNAL PITLOT STUDY			5. FUNDING NUMBERS	
6. AUTHOR(S) MAJ WEBB TIMOTHY S				
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) UNIVERSITY OF COLORADO DENVER HSC			8. PERFORMING ORGANIZATION REPORT NUMBER CI02-851	
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES) THE DEPARTMENT OF THE AIR FORCE AFIT/CIA, BLDG 125 2950 P STREET WPAFB OH 45433			10. SPONSORING/MONITORING AGENCY REPORT NUMBER	
11. SUPPLEMENTARY NOTES				
12a. DISTRIBUTION AVAILABILITY STATEMENT Unlimited distribution In Accordance With AFI 35-205/AFIT Sup 1			12b. DISTRIBUTION CODE	
13. ABSTRACT (Maximum 200 words)				
<div style="text-align: center;"> <p>DISTRIBUTION STATEMENT A Approved for Public Release Distribution Unlimited</p> <p style="font-size: 2em; font-weight: bold;">20030221 177</p> </div>				
14. SUBJECT TERMS			15. NUMBER OF PAGES 102	
			16. PRICE CODE	
17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT	

**EXACT TEST SIZE AND POWER FOR SMALL SAMPLES USING AN
INTERNAL PILOT STUDY**

by

TIMOTHY S. WEBB

B.S., United States Air Force Academy, 1988

M.S., Air Force Institute of Technology, 1994

**A thesis submitted to the
Faculty of the Graduate School of the
University of Colorado in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
Biometrics Program**

2003

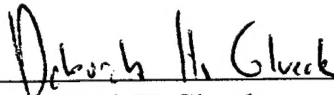
This thesis for the Doctor of Philosophy degree by

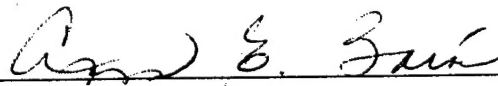
Timothy S. Webb

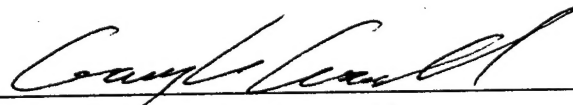
has been approved for the

Biometrics Program

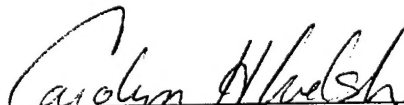
by


Deborah H. Glueck


Anna E. Barón


Gary Grunwald


John Kittelson


Carolyn Welsh

Date December 1, '02

Webb, Timothy S. (Ph.D., Analytic Health Sciences - Biometrics)

Test Size and Power for a Case-Control Study Using Internal Pilot Study Data

Thesis directed by Associate Professor Anna E. Barón.

Marlar and Welsh (2001) proposed a prospective ascertainment strategy for a case-control study of deep venous thrombosis with a binary exposure. This study design relied on prevalence estimates of certain genetic polymorphisms from previous studies using different populations. They planned on collecting a portion of the data to use as an internal pilot in order to better estimate these prevalences using their population of interest. The internal pilot data were used to estimate the population prevalences and recalculate the required sample size. All the data, including the internal pilot data, were used in the final analysis. They used a χ^2 to analyze the resulting 2×2 table. We have been unable to find a paper in which the authors considered power and sample size for a case-control study with an internal pilot study.

We solved the problem substantially. We find the distribution of the test statistic for an internal pilot study design by enumerating every 2×2 table. This design induces a dependence of the test statistic on the internal pilot data. We prove a general theorem about power for categorical tests. We derive the exact small sample power and test size for the conditional test, which conditions on both margins. We derive the exact small sample power and test size for the unconditional test, which conditions on only the number of cases and controls. We describe and, when necessary, provide methods to control the inflation in the Type I error rate caused by using an internal pilot. For the conditional test in small samples, we get the benefit of a better design without paying a penalty in terms of Type I error rate inflation above the nominal. However, this is not true for the unconditional test in which the Type I error rate is inflated even in small samples. An example is provided using data from the deep

venous thrombosis study. This work provides guidance to the practitioner interested in using an internal pilot study to re-estimate the population parameters necessary for recalculating the sample size for a binary exposure. This work provides a starting point for future research.

The form and content of this abstract are approved. I recommend its publication.

Signed Arnd S. Zahr
Faculty member in charge of thesis

The views expressed in this document are those of the author and do not reflect the official policy or position of the United States Air Force, Department of Defense, or the U.S. Government.

ACKNOWLEDGMENTS

First of all, I would like to thank my wife and children for allowing me the opportunity to go back to school, once again. As my wife will tell you, I'm not that easy to live with while in student mode. I certainly could not have made it without their love and support.

Next, I would like to thank the faculty and staff in the department of Preventive Medicine and Biometrics for getting me through the program. It has truly been my honor to be associated with such an excellent group of people. Also, I would like to thank my fellow students, especially Melanie and David, I would never have made it without their help and support.

Finally, I would like to express my appreciation to my dissertation committee. I have never worked so hard and learned so much. I am especially grateful for the support and guidance from Anna and Deb - they believed in me and never let me give up. And to Abe, the youngest member of my dissertation committee, thanks for keeping everything in perspective.

TABLE OF CONTENTS

CHAPTER

I. INTRODUCTION	1
1.1 Problem	1
1.2 Literature Review	7
1.3 Definitions and Notation	16
II. THE EXACT CONDITIONAL TEST	31
2.1 Sampling Scheme/Probability Structure	32
2.2 Exact Small Sample Power for the Conditional Test	37
2.3 Summary and Conclusions	55
III. THE EXACT UNCONDITIONAL TEST	56
3.1 Sampling Scheme/Probability Structure	57
3.2 Exact Small Sample Power for the Unconditional Test	59
3.3 Summary and Conclusions	68
IV. MOTIVATING EXAMPLE	70
4.1 Exact Conditional Test	72
4.2 Exact Unconditional Test	77
4.3 Summary and Conclusions	80
V. DISCUSSION	81
5.1 Summary and Conclusions	81
5.2 Directions for Future Research	83
BIBLIOGRAPHY	85

APPENDIX	90
A. Definitions and Notation	90
B. More Complete Table of Significance Levels	91

LIST OF TABLES

TABLE

1.	All Possible Margins and Marginal Probabilities given Internal Pilot Study Sample Size, $N_P(\cdot, \cdot)$	34
2.	Marginal Probabilities given Internal Pilot Study Sample Size, $N_P(\cdot, \cdot)$	35
3.	Example of Adjusted Product Binomial Probabilities for $\pi_1 = \pi_2 = 0.1$	38
4.	Example of Multiple Test Statistic Values	40
5.	Test Size Inflation with an Internal Pilot Study, Sample Size = 24, 48, 96	47
6.	Example of Adjustment Procedure at the Null, $\pi_1 = \pi_2 = 0.5$	51
7.	Values of Nuisance Parameter, π , for Initial Sample Sizes of 24, 48, and 96	64
8.	Test Size Inflation with an IPS, Initial Sample Size = 24, 48, 96	66
9.	Example of Adjustment Procedure at the Null, $\pi_1 = \pi_2 = \pi$	67
10.	Results from Improved Sample Size Methodology	77

LIST OF FIGURES

FIGURE

1.	2×2 Table for Initial Data	18
2.	2×2 Table for Internal Pilot Study Data	20
3.	2×2 Table for Final Data	22
4.	2×2 Table for Additional Data	23
5.	2×2 Table for Combined Internal Pilot Study and Additional Data	23
6.	Possible 2×2 Internal Pilot Study Tables	24
7.	Enumerated Additional 2×2 Tables Based on $\text{Table}_P(1)$	25
8.	Enumerated Final 2×2 Tables Based on $\text{Table}_P(1)$	26
9.	Sample Sizes from Two Different Internal Pilot Study Tables	28
10.	Example of 2×2 Tables with same Test Statistic Value	32
11.	Critical Values Without an Internal Pilot Study, $N = 24$	44
12.	Discrete Significance Levels Without an Internal Pilot Study, $N = 24$	44
13.	Example Power Plots for Initial Sample Size of 24	45
14.	Example Power Plots for Initial Sample Size of 48	46
15.	Example Power Plots for Initial Sample Size of 96	46
16.	Test Size Inflation	48
17.	Where Test Size Adjustment is Required	50
18.	Iterative Sample Size Approach	54
19.	Test Size Values for Initial Sample Sizes = 24, 48, 96	63
20.	Power Curve for $N = 24, 48$, and 96	65

21.	2×2 Contingency Table of Available Marlar and Welsh Data	70
22.	Power Curve for $N_I(\cdot, \cdot) = 93$, Conditional Approach	73
23.	Internal Pilot Study 2×2 Table from Random Experiment	74
24.	Additional and Final 2×2 Tables from Random Experiment	75
25.	Internal Pilot Study 2×2 Table from Random Experiment	79
26.	Additional and Final 2×2 Tables from Random Experiment	79

CHAPTER I

INTRODUCTION

1.1 Problem

In many situations, investigators are unsure of their parameter estimates at the beginning of a study. These parameter estimates, such as for the variance or proportion, are used to calculate the required sample size to ensure a meaningful study. It is extremely important to have the correct sample size as they do not want to have either an over-powered or under-powered study. The investigators may have parameter estimates from a different population or similar factors, but they are concerned about using these estimates since they are clearly incorrect and will lead to an incorrect sample size.

What should the investigators do at this point? One option would be to consider a pilot study. The investigators could then get estimates based on the specific population or factors of interest. An external pilot study is a very common procedure that has been used for many years. A recent advance in the literature has been the use of an internal pilot study. The main difference between the two is that in an external pilot study, the data is discarded after the parameters of interest have been estimated and this is not the case with an internal pilot study. In an internal pilot study, the pilot data is used as part of the final analysis. The advantages to using an internal pilot study are that even with a relatively large internal pilot study sample size, the investigators are not as concerned with the time and cost, since the data will not be discarded.

The problem, however, is that there is now a dependence of the final data, and hence final test statistic, on the internal pilot study data. The problem has been resolved in the continuous normal case, but not in the small sample binary case, which is the

focus of this thesis. Although the normal solutions apply even for binary outcomes as long as the sample size is large enough, we focus mainly on the small sample case and cannot apply these normal solutions. This thesis answers this problem - how do we account for this dependence of the final data on the internal pilot study data?

Specifically, we will be examining data from an observational study with a binary outcome and exposure. The goal will be to calculate the sample size and power for an observational study with an internal pilot. The data will be used to re-estimate the required final sample size. The question of interest is "Do cases have more deleterious genetic polymorphisms than the control group?" The null hypothesis to be tested is that there is an equal proportion of cases and controls with deleterious genetic polymorphisms. In the process of re-estimation, the test size for the final analysis will be inflated (Birkett and Day, 1994; Browne, 1995; Coffey and Muller, 1999 and 2001; Day, 2000; Denne, 2001; Gould, 2001; Herson and Wittes, 1993; Jennison and Turnbull, Chapter 14, 2000; Wittes and Brittain, 1990; Wittes *et al.*, 1999; Zucker *et al.*, 1999). If the inflation goes above the nominal, we need to compensate for the inflation of the test size caused from peeking at the data. This requires us to know something about the distribution of the test statistic based on the interim sample size recalculation. No current theory exists, so we develop it in this thesis.

1.1.1 Motivation

This project is motivated by the case-control study "Epidemiology of genetic and acquired risk factors for venous thrombosis" (Marlar and Welsh, 2001). Venous thromboembolism is a very common medical problem with high morbidity and a significant mortality. Venous thromboembolism is an obstruction in a vein caused by a dislodged thrombus or clot. If a person with high risk of venous thromboembolism is identified, the complications (such as death) of this disease or condition can be minimized. Risk factors for venous thromboembolism are incompletely understood

although both genetic and acquired factors are involved. Due to the multifactorial nature of venous thromboembolism (Marlar, 2001), these risk factors undoubtedly interact to cause venous thromboembolism for a specific person. Identified high risk acquired risk factors for venous thromboembolism include surgery, trauma, immobility, antiphospholipid antibodies, and malignancy.

Recently, several single gene polymorphisms have been described which themselves increase risk for venous thromboembolism. The impact of inheriting two or more prothombotic genetic polymorphisms on the development of venous thromboembolism is not well understood although it does appear that a person's risk of venous thromboembolism is significantly increased with the inheritance of more than one of these genetic polymorphisms. In addition, the impact of several of these inherited risk factors in combination with high risk acquired conditions, such as surgery, has not been studied.

Certain genetic polymorphisms have been found to increase the risk of venous thromboembolism both for those with and without (identifiable) acquired risk conditions. Marlar and Welsh (2001) are trying to determine whether the inheritance of any of the following polymorphisms is associated with a higher risk of venous thromboembolism: factor V Leiden, homocysteine-MTHFR, prothrombin-20210, PLA2, the ACE D/I deletion insertion polymorphism, and factor VIII levels.

The study is planned for a ratio of cases to controls of one to two. Cases and controls are matched on gender and age (\pm 5 years). The original sample size planned is 900 subjects, 300 cases and 600 controls. The study is currently more than half way through recruitment: there are approximately 345 controls and 186 cases. The design results in a minimally detectable odds ratio of 1.7 - 3.6, depending on the prevalence of the polymorphism, at a power of 90%. From the literature, the prevalence of the polymorphisms of interest ranged from as high as 10% to as low as 1%. The minimally

detectable odds ratio for this project is based on assumptions for a study without an internal pilot and may need adjustment if an internal pilot is used. The advantage of an internal pilot is that estimates are based on the population being studied. Since many of the venous thromboembolism information is from Northern European populations, this may give an over or underestimate for the genetically more heterogeneous United States population.

1.1.2 Study Design

Before we can solve the problem, let us carefully describe the process by which the internal pilot study is performed. The first step is to estimate the initial sample size. The initial sample size is based on a minimally detectable difference (or odds ratio) and target power. Currently, we are using the sample size equation proposed by Rosner (1995) for comparing two binomial proportions. We realize that this equation is inappropriate to use in small samples because it is based on the asymptotic distribution of the difference in two proportions. However, we plan on using it at this point as a means to simplify the problem and address the main concern of accounting for the dependence of the final data on the internal pilot study data. The exact sample size formula will be provided in § 2.2.6 and § 3.2.6; but first, we must develop some notation and derive the distribution of the final sample size and hence the final test statistic. Note the maximum sample size might be dictated by either time, money, or population size. Additionally, we have assumed a ratio of cases to controls of one to two.

Once the initial sample size has been calculated, a percentage of it, such as 50%, is used as the internal pilot study. What this means is that 50% of the initial sample size is collected as the internal pilot study; for example if the initial sample size is 60, then the internal pilot study sample size would be 30 (10 cases and 20 controls). Once the internal pilot study data have been collected, we can then re-estimate the parameters

of interest, in this case, the proportion of cases and controls with the genetic polymorphism of interest. These new parameter estimates are then used to re-estimate the final sample size based on the re-estimated difference in proportions and target power. Once again, we are using the sample size equation proposed by Rosner (1995) for simplicity. If the re-estimated final sample size, for example, is 90, then we would need to collect an additional sample of size 60, or 20 cases and 40 controls. Since we are interested in small samples, we are limiting the re-estimated final sample size to twice that initially estimated. Depending on what the investigators are interested in, the sample size calculation could change. In this case, we are looking specifically at the difference of proportions; however, the odds ratio or the ratio of proportions might also be of interest. The sample size and sample size re-estimation procedures would change accordingly.

Once the additional samples have been collected, the analysis is performed on all of the data. What that means is that we combine the internal pilot study data with the additional samples and then perform the data analysis on the combined data as if it were a single data set. However, we must develop a methodology to account for the dependence of the additional sample size on the internal pilot study data.

1.1.3 Specific Aims

The goal of this dissertation is to accurately calculate sample size and power for a binary outcome case-control study with an internal pilot. Initial sample size estimates for the motivating study are based on the prevalence of certain genetic polymorphisms from previous studies and possibly other populations of interest. The population that Marlar and Welsh (2001) are investigating are patients at the Veteran's Administration and University Hospitals. The prevalence of the genetic polymorphisms of interest may not be the same as those from the previous studies on which the initial sample sizes are based.

For this thesis, we develop two different inferential approaches. Both approaches use an internal pilot to re-estimate the parameters of interest in order to recalculate the necessary sample size to ensure an adequately powered study. The first approach, Chapter II, is the exact conditional one and the second approach, Chapter III, is the exact unconditional one. Conceptually, the main difference is that the conditional approach conditions not only on the sample size but also on the number of cases and controls with the genetic polymorphism of interest and the unconditional approach only conditions on the sample size. We examine test size inflation using these two approaches after including an internal pilot study. We also provide a method or procedure to account for the test size inflation caused by using an internal pilot study. Finally, we provide a simple example from Marlar and Welsh (2001) in Chapter IV and possible extensions and enhancements to this thesis in Chapter V.

1.1.4 Settings and Assumptions

We have several settings and simplifying assumptions. First, the outcome of interest (disease or no disease) and predictor (genetic polymorphism or not) are both binary categorical variables. The population parameters of interest will be varied across a range to examine the behavior of these methods at different values. We are using the internal pilot study to re-estimate these parameters of interest for the sample size recalculation. We have assumed no confounders or interactions at this point. Therefore, this simplified situation can be analyzed using a 2×2 table. We will be focusing on the difference in proportions.

Additionally, we placed some restrictions on the sample size. We will never let the final sample size be smaller than that initially estimated or larger than twice that initially estimated. We believe these are reasonable limits to place on the sample size.

Since we are using an internal pilot study, we are essentially peeking at the data even though no hypothesis tests are carried out. Just as in an analysis for sequential

designs, a penalty must be paid for this peeking. The result of using the revised parameter estimates from the internal pilot study to inform the revised sample size calculation is an inflated test size with respect to the discrete significance level and possibly the nominal significance level. Therefore, we need to account for the inflated test size in determining the significance (based on either the p -value or critical value) of the final test statistic applied to the combined internal pilot and additional data. The distribution of the final test statistic is complicated by the dependence of the additional sample size on the observed internal pilot study.

1.2 Literature Review

We have been unable to find a paper in which the authors considered power and sample size for a case-control study with an internal pilot study. Papers, however, have been written on power and sample size for the continuous dependent variable case with an internal pilot study and also power and sample size for categorical data without an internal pilot study. These papers form the background and inspiration for this work. We give a short review of the minor topics, followed by a longer description of the underpinnings of this work.

Next, we will discuss some issues relevant to sample size re-estimation and internal pilot studies.

1.2.1 Tests for the 2×2 Tables

There are several different classes of tests for 2×2 tables. In the next few sections, we will carefully review the differences between these tests.

1.2.1.1 Asymptotic versus Exact

In our motivating example (Marlar and Welsh, 2001), there is a binary outcome and a binary exposure. As part of our assumptions, we will focus on the difference in proportions of the exposure between the two outcome groups. Part of the reason for

this is that the parameter estimates of the proportion of cases and controls with the genetic polymorphism are of interest.

Yates (1984) and Little (1989) provide background information on comparing two binomial proportions or testing the significance for 2×2 contingency tables. Both of these authors examine Fisher's exact, Pearson's χ^2 , and Yates's continuity-corrected χ^2 tests. Asymptotically, these three methods are equivalent. In small to moderate samples Fisher's exact and Yates's continuity-corrected χ^2 tests are more conservative than Pearson's χ^2 test. By more conservative, we mean that the p -values for the Fisher's exact and Yates's continuity-corrected χ^2 tests are larger than the corresponding Pearson's χ^2 test.

Satten and Kupper (1990) and Greenland (1988), however, suggest using interval estimation for significance testing which they state is more appropriate in the epidemiological literature. Consequently, they also suggest using interval estimation for their sample size requirements. Satten and Kupper suggest the use of the odds ratio while Greenland works more generally with 2×2 contingency tables. He uses confidence intervals that are equivalent to or based on significance testing.

1.2.1.2 Conditional versus Unconditional Hypothesis Tests

With case-control data the standard methods are based on conditional inference. We could, however, use either conditional or unconditional inference. The conditional approach assumes both sets of marginals are fixed as in Fisher's exact test with the inference also based on fixing both margins. The unconditional approach, however, fixes only one margin, the number of cases and controls which also dramatically changes the inference. The unconditional inference applies to all possible numbers of cases and controls with the genetic polymorphism of interest. The conditional inference fixes the number of cases and controls with the genetic polymorphism of interest. The unconditional approach was suggested by Barnard (1947).

There has been quite a bit of debate about whether to use a conditional or unconditional test. Even Barnard (1949), the father of the unconditional test, later turned against it. Little (1989) states that "Yates and others (1984) argued that both [margins] should be held fixed for inference, even though only one margin is fixed by the product-binomial sampling design." Little also states that using the conditional approach (that is conditioning on both margins) "makes the test more conservative" than the unconditional approach because there are more tables possible with the unconditional approach.

Little (1989), Yates (1984), and others believe that it is inappropriate to use the unconditional approach. Specifically, Little suggests that if the inference on the odds ratio changes dramatically from moving from the conditional to the unconditional approach, then the unconditional approach is suspect. Suissa and Shuster suggest that the strong motivation for using the unconditional approach is "the ease of explanation of the results." In the next section, we will review the literature for an internal pilot study, the analysis of which could rely on either a conditional or unconditional approach.

1.2.2 Internal Pilot Studies

Appropriate estimation of sample size is key to the success of any clinical trial. In order to estimate the sample size appropriately, accurate estimates are required for the treatment effect and variability. In many cases, this is not possible due to the study of new therapies and unfamiliar diseases. By misspecifying either the treatment effect or variability, the sample size could either be too large or too small, which results in either a non-economical or an under-powered study. An important aspect of an internal pilot is the ability to use data from the current study to help get a better estimate of these parameters, e.g. the variance, and hence the required sample size.

Case-control studies typically examine the possible relationship between an exposure and disease (Gordis, 1996). In this situation, we are examining the possible relationship between one or more genetic polymorphisms and venous thromboembolism (Marlar and Welsh, 2001). Subjects with venous thromboembolism are cases and those without venous thromboembolism are controls. Although this is an observational study, the justification of sample size is just as important as in a clinical trial. We do not want to miss an important effect due to an inadequately sized study. In this example, if a genetic polymorphism is associated with venous thromboembolism, we would expect the proportion of cases with the genetic polymorphism to be greater than the proportion of controls without the genetic polymorphism.

Internal pilot studies have been seen in the clinical trials literature since 1990. The reason for using pilot studies is the uncertainty around certain parameter estimates used in calculating the required sample size. In this example, we are uncertain of the proportion of cases and controls with the genetic polymorphism of interest since initial estimates are based on previous studies from different populations. Thus, the original sample size calculation is almost surely wrong. One of the advantages of using an internal pilot study is that we know for certain that the new estimates are from the correct population. Another advantage of using an internal pilot study versus an external one is that the data collected as part of the internal pilot study are not lost, i.e., the data from the internal pilot study are used in the final analysis.

Wittes and Brittain (1990) define two types of pilot studies: 1) "external" pilot - "a study that is structurally distinct from the main study" and 2) "internal" pilot - a study that "is an integral part of the main investigation." They argue that the use of each type of pilot study depends on the purpose of the study. For example, external pilot studies should be used to determine if the process is feasible (i.e., the mechanics of the trial), while internal pilot studies should be used when the process appears feasible but there

is some uncertainty associated with the study parameters. They provide an example involving a normally distributed outcome variable for ease of computation. The ideas presented are equally applicable to other types of outcome variables, such as binary, but "can be more difficult or challenging" according to Wittes and Brittain. Also, unlike external pilots, internal pilots can be large with essentially no effect on study length or cost. Wittes and Brittain are one of the first to use the term internal pilot study. In their paper, they propose to use a percentage of the main trial as a pilot phase. This is why they used the term "internal" pilot study. The purpose for doing this is to better determine or estimate the values of unknown parameters, such as the variance or event rates. With these new estimates, they propose to recalculate the required sample size to ensure an adequately powered study. The final analysis is based on all of the data.

Wittes and Brittain (1990) state that investigators "should have reliable prior estimates of three classes of parameters, related to (1) the administration of the study, (2) the process of the disease and (3) the effect of treatment." They view variables or parameters related to the "administration of the study" and "process of the disease" as candidates for either internal or external pilots. Parameters related to "the effect of treatment," they regard as appropriate for external pilots only. Their interest lies in the parameters that relate to the process of the disease.

Wittes and Brittain provide an example of a normally distributed outcome variable in which the variance parameter is of interest. Before beginning a trial to compare two groups (treatment and control), an estimate of the variance is needed. Based on this variance, a required sample size can be calculated. For their example, they use half of the originally calculated sample size for an internal pilot study to re-estimate the variance parameter. Based on the variance re-estimation, the required sample size is recalculated. If the recalculated sample size is less than or equal to the originally planned size, the study is continued as originally planned. However, if the

recalculated sample size is greater than the originally planned size, the study is continued using the recalculated sample size.

Wittes and Brittain note that the Type I error, α , can be inflated by this procedure with the greatest effect produced when the re-estimated variance is closest to the original variance estimate. However, they state that in most practical situations this inflation is only slight. The design used by Wittes and Brittain is a variant of Stein's (1945) classic two-stage procedure. Stein's two-stage procedure essentially uses what we call the internal pilot to re-estimate the variance but does not use these same observations in the final variance calculation. There are a few significant differences between what Wittes and Brittain do as compared with Stein's two-stage procedure. Wittes and Brittain do not allow the sample size to be smaller than the originally estimated one. They also use all the data to re-estimate the variance instead of only the data from the first stage (pilot) as done by Stein. In conclusion, Wittes and Brittain believe that internal pilots are very useful when the variance is understated which they believe is typical of clinical trials. Additionally, in cases such as these, the effect on the observed test size is minimal and the chances of having a meaningful study are greatly increased.

Wittes and Brittain (1990) and Shih and Zhao (1997) consider internal pilot studies for the re-estimation of sample size. These authors note that using data from an internal pilot study to re-estimate the sample size has an effect on test size. Coffey and Muller (1999) assume the initial parameter estimates are fixed and known and address the issue of using an internal pilot study to re-estimate the parameters of interest and using these to revise the required sample size. Coffey and Muller do this in the context of the general linear univariate model with the simplest case being a t -test. By using an internal pilot study to re-estimate the sample size, the test size is inflated and this must

be taken into account in a subsequent analysis. Coffey and Muller are able to account for this test size inflation in their analysis.

Coffey and Muller (1999) state that the power of a test depends on one or more unknown nuisance parameters, such as the error variance. Therefore, the required sample size for a specified power depends on this variance estimate. They provide five steps in testing a hypothesis for a general linear univariate model, with Gaussian errors, similar to that suggested by Wittes and Brittain (1990). Coffey and Muller describe methods for computing exact test size and power under this general linear univariate model scenario.

In order to estimate this variance, we could use an educated guess or an estimate from a prior study. Using a prior study or external pilot study for estimating the variance has the advantage that the observations used in the variance estimate are independent of those in the planned analysis. However, the observations from the external pilot study may be very different from those for the planned analysis. By using an estimate of the variance, randomness is introduced into the sample size value. However, using estimates from prior studies may lead to great uncertainty in the variance estimate due to the possibly different population under study and the possibly more homogeneous population than used in later trials. Therefore, using an internal pilot study is very appealing since the variance estimate is actually based on data from the current study.

Unfortunately, the observations used in the variance estimate are no longer independent of those used in the final analysis because the test statistic now depends on the variance estimate through the final sample size. Wittes and Brittain state that the effect on the test size is minimal for moderately large sample sizes when used with the two-sample t -test and not allowing a decrease in initial sample size calculations. Coffey and Muller were motivated by more complex designs and smaller sample sizes. They

desire to compute the exact distribution of the test statistic so the effect or bias on the test size can be controlled and therefore used for smaller sample sizes or more complex designs. Coffey and Muller derived the exact distribution of the test statistic while using an internal pilot study for comparing two groups with a general linear model. However, Coffey and Muller state that their results are more general with the added benefit of allowing sample sizes to decrease from those initially estimated if desired.

While no author has specifically considered internal pilot studies for categorical data, there is extensive literature on sample size for conditional and unconditional tests with categorical data. There are sample size estimates for both the conditional and unconditional tests. In the conditional case, sample size and power estimates are based either on an asymptotic test (such as Pearson's χ^2) or the Fisher's exact test (Lachin, 1977 and 1981; Law, 1996; Little, 1989; Rosner, 1985; Schlesselman, 1974 and 1982; and Yates, 1984). With respect to sample size and power estimates in the unconditional case, there is much less literature available and even fewer options (Barnard, 1947; Little, 1989; and Suissa and Shuster, 1985). Suissa and Shuster (1985) derive exact sample sizes using an unconditional Z statistic in 2×2 contingency tables for two independent binomial samples of equal size using the approach suggested by Barnard. Defining π_1 and π_2 as the proportion of controls and cases, respectively, with the genetic polymorphism of interest, what Suissa and Shuster propose is an iterative approach under the null hypothesis, $H_0: \pi_1 = \pi_2 = \pi$, that searches across the proportion of cases and controls, now equal to π , until the maximum observed significance level less than or equal to the nominal significance level is found. This maximum observed significance level is the test size. By doing this, they have essentially removed from the problem the nuisance parameter, π . These unconditional statistics produce required sample sizes smaller than their exact conditional counterparts. They state that this is the first paper in which sample size calculations are

based on exact unconditional tests for which no ancillary statistic exists for the nuisance parameter. However, they do not consider sample size re-estimation using an internal pilot study.

1.2.3 Blinded versus Unblinded Re-estimation of Sample Size

One method of controlling test size inflation is to remain blinded at the internal pilot stage. If maintaining the blind is possible, this would be the preferred method, especially by regulatory agencies (Gould, 2001). By maintaining the blinding, there is much less chance of introducing any sort of bias even if unintentional. Shih and Zhao (1997) recommend not unblinding to help control the test size inflation. They note, however, that there is still a moderate inflation in test size while maintaining the blind, but the inflation is not as large as procedures that unblind the data. Gould (1992) and Shih and Zhao present an approach designed for binary response variables which is more difficult than the normal case because the variance is not independent of the mean. These articles provide a method to re-estimate sample size while remaining blinded. Herson and Wittes (1993) also considered the case of a binary response, however, they did not remain blinded. Shih and Zhao state that both of these articles are not adequate because the treatment effect must remain unknown. For the observational study we are considering here (Marlar and Welsh, 2001), remaining blinded at the sample size re-estimation stage is not an option and cannot affect the outcome. In a case-control study, the manner in which subjects are chosen, e.g. a case or a control, prohibits blinding.

1.2.4 Summary of the Literature

As outlined in the above sections, there has been a great deal of work done in the area of sample size re-estimation using internal pilot studies. However, the majority of the work has been in the area of continuous outcome such as those considered by Wittes and Brittain (1990) and Coffey and Muller (1999). Specifically, Coffey and

Muller are able to compute the exact distribution of the final sample size and hence of the final test statistic for the general linear univariate model. Therefore, they are able to control for the inflation in test size. In terms of binary data, there has been some work done by Gould (1992), Herson and Wittes (1993), Shih and Zhao (1997), and Gould (2001). However, the distribution of the final test statistic nor any adjustment to either the test statistic or nominal significance level to account for the inflated test size has been reported in the current literature. One method proposed to account for this test size inflation is to maintain the blinding (Shih and Zhao, 1997). However, this is not possible with the data from Marlar and Welsh (2001) because of the nature of case-control data. Additionally, there has not been any literature regarding the use of an unconditional approach with an internal pilot study.

After reviewing the current literature, it is clear that the problem of sample size re-estimation using binary data has not been done. Next, we will provide the definitions and notation necessary for understanding the remainder of this thesis.

1.3 Definitions and Notation

In this section, the data structure and notation are presented for the initial, internal pilot, and final phases of a case-control study with binary outcome and exposure. See Appendix A for a full listing of the notation used throughout this thesis.

Let the ratio variable, k , be the ratio of controls to cases. Given the proportion of cases and controls with a specific genetic polymorphism, $p(1)$ and $p(2)$ respectively, we can define the following variables;

$$q(1) = 1 - p(1), \tag{1}$$

$$q(2) = 1 - p(2), \tag{2}$$

$$\bar{p} = \frac{p(1) + kp(2)}{1 + k}, \quad (3)$$

and

$$\bar{q} = 1 - \bar{p}. \quad (4)$$

We add subscripts to these variables to indicate whether they are initial (I), internal pilot (P), additional (A), or final (F). For example, let p_I , p_P , and p_F indicate the proportion of the initial, internal pilot, and final sample sizes with the genetic polymorphism of interest. Let q_I , q_P , and q_F be defined as $(1 - p_I)$, $(1 - p_P)$, and $(1 - p_F)$, respectively.

Further, let $Z_q = q^{\text{th}}$ quantile of a standard normal distribution. Finally, let the probability of rejecting the null hypothesis, when the null is true, be defined as the nominal significance level, α . For a specified alternative, let the probability of rejecting the null be defined as the power, $(1 - \beta)$.

1.3.1 Initial Sample Size

In this section, we will introduce the notation and calculations for the initial sample size. The initial sample size is based on the assumption that the proportion of cases and controls with the genetic polymorphism of interest are fixed and known. In actual practice, this assumption may not be true since usually the proportions with the genetic polymorphism of interest are estimated from previous studies. Let $n_I(1, 1)$ be the number of cases with the genetic polymorphism of interest. Let $n_I(1, 2)$ be the number of cases without the genetic polymorphism of interest. Define $n_I(2, 1)$ and $n_I(2, 2)$ as the number of controls with and without the genetic polymorphism of interest, respectively. Similarly, we can define $n_P(1, 1)$, $n_P(1, 2)$, $n_P(2, 1)$, $n_P(2, 2)$, $n_A(1, 1)$, $n_A(1, 2)$, $n_A(2, 1)$, $n_A(2, 2)$, $n_F(1, 1)$, $n_F(1, 2)$, $n_F(2, 1)$ and $n_F(2, 2)$ for

the internal pilot, additional, and final data. Define the marginal totals and the overall total as shown in Figure 1.

	Genetic Marker		
	+	-	
Cases			$N_I(1, \cdot)$
Controls			$N_I(2, \cdot)$
	$N_I(\cdot, 1)$	$N_I(\cdot, 2)$	$N_I(\cdot, \cdot)$

Figure 1. 2×2 Table for Initial Data.

Let \bar{p}_I be defined as

$$\bar{p}_I = \frac{p_I(1) + k \cdot p_I(2)}{1 + k}, \quad (5)$$

and \bar{q}_I be defined as

$$\bar{q}_I = 1 - \bar{p}_I. \quad (6)$$

We can now give the equation for calculating the initial sample size for the cases, $N_I(1, \cdot)$, for comparing two binomial proportions (Rosner, 1995) as

$$N_I(1, \cdot) = \frac{\left[\sqrt{\bar{p}_I \bar{q}_I \left(1 + \frac{1}{k}\right)} Z_{1-\alpha/2} + \sqrt{p_I(1)q_I(1) + \frac{p_I(2)q_I(2)}{k}} Z_{1-\beta} \right]^2}{[p_I(1) - p_I(2)]^2} \quad (7)$$

The initial sample size for controls, $N_I(2, \cdot)$, can then be calculated using the ratio of controls to cases, k ,

$$N_I(2, \cdot) = k \cdot N_I(1, \cdot). \quad (8)$$

The sum of these

$$N_I(\cdot, \cdot) = N_I(1, \cdot) + N_I(2, \cdot) \quad (9)$$

is the total initial sample size.

1.3.2 Internal Pilot Study Sample Size

In this section, we will introduce the notation and calculations for the internal pilot study sample size. Once the initial sample size has been calculated, the internal pilot study sample size can be calculated. The internal pilot study sample size is usually based on some percentage, " a ", of the initial sample size. After deciding on the percentage of the initial sample size, the size of the internal pilot study for the cases and controls, respectively, can be calculated by

$$N_P(1, \cdot) = \frac{aN_I(\cdot, \cdot)}{(k + 1)} \quad (10)$$

and

$$N_P(2, \cdot) = kN_P(1, \cdot) \quad (11)$$

and rounding up to the nearest integer or subject. After calculating $N_P(1, \cdot)$ and $N_P(2, \cdot)$, we can then calculate the total internal pilot study sample size by adding the number of cases and controls,

$$N_P(\cdot, \cdot) = N_P(1, \cdot) + N_P(2, \cdot). \quad (12)$$

Once the data have been collected, we can create the 2×2 table for the internal pilot study data as shown in Figure 2.

	Genetic Marker		
	+	-	
Cases	$n_P(1, 1)$	$n_P(1, 2)$	$N_P(1, \cdot)$
Controls	$n_P(2, 1)$	$n_P(2, 2)$	$N_P(2, \cdot)$
	$N_P(\cdot, 1)$	$N_P(\cdot, 2)$	$N_P(\cdot, \cdot)$

Figure 2. 2×2 Table for Internal Pilot Study Data.

With these data, we can re-estimate π_1 and π_2 , the population proportion of cases and controls, respectively, with the genetic marker of interest. These new estimates of π_1 and π_2 , $p_P(1)$ and $p_P(2)$ respectively, are calculated using the data from the 2×2 table as follows:

$$p_P(1) = \frac{n_P(1, 1)}{N_P(1, \cdot)} \quad (13)$$

and

$$p_P(2) = \frac{n_P(2, 1)}{N_P(2, \cdot)}. \quad (14)$$

1.3.3 Additional/Final Sample Size

In this section, we will introduce the notation and calculations for the additional and final sample sizes using the new estimates, $p_P(1)$ and $p_P(2)$, calculated above. As defined for the initial sample size, we can similarly define

$$\bar{p}_P = \frac{p_P(1) + k \cdot p_P(2)}{1 + k}, \quad (15)$$

and

$$\bar{q}_P = 1 - \bar{p}_P. \quad (16)$$

The final sample size, $N_F(\cdot, \cdot)$, can be calculated using the sample size equation to compare two binomial proportions (Rosner, 1995) once again. This equation using our notation is

$$N_F(1, \cdot) = \frac{\left[\sqrt{\bar{p}_P \bar{q}_P \left(1 + \frac{1}{k}\right)} Z_{1-\alpha/2} + \sqrt{p_P(1)q_P(1) + \frac{p_P(2)q_P(2)}{k}} Z_{1-\beta} \right]^2}{[p_P(1) - p_P(2)]^2} \quad (17)$$

where the "1" stands for cases and the "." notation stands for the total number of cases. The final sample size for controls, $N_F(2, \cdot)$, where "2" stands for the controls and "." stands for the total number of controls, can then be calculated using the ratio of controls to cases, k ,

$$N_F(2, \cdot) = k N_F(1, \cdot). \quad (18)$$

The sum of these

$$N_F(\cdot, \cdot) = N_F(1, \cdot) + N_F(2, \cdot) \quad (19)$$

is the total final sample size. This is not the number of additional subjects needed, but the total number of subjects. The 2×2 table for the final data, after collecting the additional data, is provided in Figure 3.

	Genetic Marker		
	+	-	
Cases	$n_F(1, 1)$	$n_F(1, 2)$	$N_F(1, \cdot)$
Controls	$n_F(2, 1)$	$n_F(2, 2)$	$N_F(2, \cdot)$
	$N_F(\cdot, 1)$	$N_F(\cdot, 2)$	$N_F(\cdot, \cdot)$

Figure 3. 2×2 Table for Final Data.

The 2×2 table for the final sample size, Figure 3, can be broken down into two 2×2 tables: one table for the internal pilot study results (Figure 2) and another table for the additional subjects used (Figure 4). The additional data required can be determined by subtracting the total internal pilot study sample size from the total final sample size,

$$N_A(\cdot, \cdot) = N_F(\cdot, \cdot) - N_P(\cdot, \cdot). \quad (20)$$

Similarly, the additional sample sizes for both the cases and controls can be calculated using

$$N_A(1, \cdot) = N_F(1, \cdot) - N_P(1, \cdot) \quad (21)$$

and

$$N_A(2, \cdot) = N_F(2, \cdot) - N_P(2, \cdot), \quad (22)$$

respectively.

	Genetic Marker		
	+	-	
Cases	$n_A(1, 1)$	$n_A(1, 2)$	$N_A(1, \cdot)$
Controls	$n_A(2, 1)$	$n_A(2, 2)$	$N_A(2, \cdot)$
	$N_A(\cdot, 1)$	$N_A(\cdot, 2)$	$N_A(\cdot, \cdot)$

Figure 4. 2×2 Table for Additional Data.

The notation from the internal pilot study and the additional data can be used to make a new 2×2 table, Figure 5, for the combined internal pilot study data and the additional data. This table provides us with the 2×2 table for the final data, Figure 3.

	Genetic Marker		
	+	-	
Cases	$n_P(1, 1) + n_A(1, 1)$	$n_P(1, 2) + n_A(1, 2)$	$N_P(1, \cdot) + N_A(1, \cdot)$
Controls	$n_P(2, 1) + n_A(2, 1)$	$n_P(2, 2) + n_A(2, 2)$	$N_P(2, \cdot) + N_A(2, \cdot)$
	$N_P(\cdot, 1) + N_A(\cdot, 1)$	$N_P(\cdot, 2) + N_A(\cdot, 2)$	$N_P(\cdot, \cdot) + N_A(\cdot, \cdot)$

Figure 5. 2×2 Table for Combined Internal Pilot Study and Additional Data.

There are several situations possible after going through this process. The final sample size calculated could be greater than, less than, or equal to the initial sample size. A decision should be made at the beginning of the study as to what will be done based on the sample size re-estimation. For example, it could be decided at the beginning of the study that the final sample size will never be smaller than the initial sample size, $N_F \geq N_I$. The only absolute requirement is that the final sample size be greater than or equal to the internal pilot study sample size, $N_F \geq N_P$, as seen by Equation 20 since we cannot have a negative sample size. Recall, we have assumed that the final sample size will never be smaller than that initially estimated, $N_F(\cdot, \cdot) \geq N_I(\cdot, \cdot)$, and the maximum sample size to be twice that initially estimated, $N_F(\cdot, \cdot) \leq 2N_I(\cdot, \cdot)$.

Before we get into defining the table probabilities and test statistic values, we must first describe the general approach we are using. The approach is based on enumeration.

1.3.4 Enumeration Approach

This thesis is based almost entirely on the enumeration process. What we mean by the enumeration process is that we examine every result, e.g. 2×2 table, possible based on a given sample size. For example, if we have calculated internal pilot study table of $N_P(\cdot, \cdot) = 3$, we know that, given this sample size and the ratio of controls to cases, $k = 2$, there are 1 case, $N_P(1, \cdot)$, and 2 controls, $N_P(2, \cdot)$. As a result, there are 6 possible internal pilot study tables, where $1 \leq i \leq 6$, Figure 6.

Table _P (1)	Marker + -	1 0 1 0 2 2 1 2 3
Table _P (2)	Marker + -	1 0 1 1 1 2 2 1 3
Table _P (3)	Marker + -	1 0 1 2 0 2 3 0 3
Table _P (4)	Marker + -	0 1 1 0 2 2 0 3 3
Table _P (5)	Marker + -	0 1 1 1 1 2 1 2 3
Table _P (6)	Marker + -	0 1 1 2 0 2 2 1 3

Figure 6. Possible 2×2 Internal Pilot Study Tables.

For each internal pilot study 2×2 table in Figure 6, we can recalculate a final sample size, $N_F(\cdot, \cdot)$, required to maintain the nominal significance level and target power level using Equations 17-19. For example, let us examine internal pilot study table 1, Table_P(1). The recalculated final sample size is $N_F(\cdot, \cdot) = 6$. Based on this recalculation, the additional sample size required $N_A(\cdot, \cdot) = N_F(\cdot, \cdot) - N_P(\cdot, \cdot) = 3$. Since our ratio of cases to controls is 1:2, our final samples sizes are $N_F(1, \cdot)$

$= 2$ and $N_F(2, \cdot) = 4$, respectively for cases and controls. Which gives us the following additional samples needed, $N_A(1, \cdot) = N_F(1, \cdot) - N_P(1, \cdot) = 1$ and $N_A(2, \cdot) = N_F(2, \cdot) - N_P(2, \cdot) = 2$. Therefore, there are 6 possible additional tables, where $1 \leq j \leq 6$. Enumerating over all 6 possible tables gives us the following additional tables, Figure 7, based on $\text{Table}_P(1)$. By combining the internal pilot study 2×2 $\text{Table}_P(1)$ from Figure 6 with the additional tables from Figure 7 we can list all possible final tables, $\text{Table}_F(1, j)$ (Figure 8). In order to enumerate over the entire sample space, we would need to similarly recalculate the final sample size required to maintain the nominal test size and target power level for each internal pilot study table, that is, over the entire set i .

Now that we have described the enumeration process which leads to every possible 2×2 table that must be analyzed, we can describe how to calculate the table probabilities.

Table_A(1, 1)

Marker

+

-

Cases

Controls

1

0

2

3

1

2

3

Table_A(1, 2)

Marker

+

-

Cases

Controls

0

1

2

2

1

1

2

3

Table_A(1, 3)

Marker

+

-

Cases

Controls

1

0

1

1

2

1

2

3

Table_A(1, 4)

Marker

+

-

Cases

Controls

0

1

1

1

2

1

2

3

Table_A(1, 5)

Marker

+

-

Cases

Controls

1

0

0

2

1

1

2

3

Table_A(1, 6)

Marker

+

-

Cases

Controls

0

1

0

2

0

1

2

3

Figure 7. Enumerated Additional 2×2 Tables Based on $\text{Table}_P(1)$.

Table_F(1, 1)

Marker

+

-

Cases

Controls

2

0

2

2

2

4

4

2

6

Table_F(1, 2)

Marker

+

-

Cases

Controls

1

1

2

2

3

3

6

Table_F(1, 3)

Marker

+

-

Cases

Controls

2

0

2

3

3

3

6

Table_F(1, 4)

Marker

+

-

Cases

Controls

1

1

2

1

3

4

2

4

6

Table_F(1, 5)

Marker

+

-

Cases

Controls

2

0

2

0

4

4

2

4

6

Table_F(1, 6)

Marker

+

-

Cases

Controls

1

1

2

0

4

4

1

5

6

Figure 8. Enumerated Final 2×2 Tables Based on Table_P(1).

1.3.5 Table Probabilities

We will need to define the probability of observing a table in order to determine the probability of rejecting the null hypothesis. Without knowing the table probabilities, we cannot determine the test size or power. There are a finite and well defined set of 2×2 tables with a given sample size and ratio of controls to cases, k . Notice that $n_P(1, 1) \in \{0, 1, \dots, N_P(1, \cdot)\}$, and similarly, $n_P(2, 1) \in \{0, 1, \dots, N_P(2, \cdot)\}$. Let B be the total number of possible internal pilot study tables. Therefore,

$$B = [N_P(1, \cdot) + 1][N_P(2, \cdot) + 1]. \quad (23)$$

Note that by enumerating the internal pilot study sample size, $N_P(\cdot, \cdot)$, we have B possible internal pilot study tables. Let $i \in \{1, 2, \dots, B\}$ index each possible internal pilot study table. Therefore, we can denote each internal pilot study table as Table_P(i) and the probability of each internal pilot study table as $\Pr\{\text{Table}_P(i)\}$.

After observing each internal pilot study table, $\text{Table}_P(i)$, we obtain a number of additional tables based on the sample size re-estimation and enumeration. Notice that each internal pilot study table may produce a different number of additional tables based on the sample size formula, Equation 17. For example see Figure 9 below which provides two possible internal pilot study tables, $\text{Table}_P(1)$ and $\text{Table}_P(2)$, that lead to different additional sample sizes, N_{A_1} and N_{A_2} . Let $N_{A_i}(\cdot, \cdot)$ be the additional sample size determined by the i^{th} internal pilot study table, such that there are $N_{A_i}(1, \cdot)$ cases and $N_{A_i}(2, \cdot) = kN_{A_i}(1, \cdot)$ controls. Therefore, we have M possible additional tables per internal pilot study table. However, we must subscript the number of additional tables with the subscript i since the M possible additional tables depends on which internal pilot study table it came from. Let M_i be the number of additional/final tables that result from the i^{th} internal pilot study table. Note that by a similar argument as above, the number of additional/final tables is defined by

$$M_i = [N_{A_i}(1, \cdot) + 1] \times [N_{A_i}(2, \cdot) + 1]. \quad (24)$$

Let j index the set of possible additional tables after recalculating the sample size based on each internal pilot study table; that is, let each additional table be denoted by $\text{Table}_A(i, j)$ where $i \in \{1, \dots, B\}$ and $j \in \{1, \dots, M_i\}$. Similarly, let each final table be denoted by $\text{Table}_F(i, j)$. Therefore, let the $\Pr\{\text{Table}_A(i, j)\}$ be the probability we observe additional table j based on internal pilot study table i . Similarly, let the $\Pr\{\text{Table}_F(i, j)\}$ be the probability we observed final table j based on internal pilot study table i .

Recall that $n_P(1, 2) = N_P(1, \cdot) - n_P(1, 1)$, $n_P(2, 2) = N_P(2, \cdot) - n_P(2, 1)$, and $2N_P(1, \cdot) = N_P(2, \cdot)$. Thus, knowing $N_P(\cdot, \cdot)$, $n_P(1, 1)$, and $n_P(2, 1)$ defines all cells and marginal totals for each internal pilot study table. Similarly,

knowing $N_F(\cdot, \cdot)$, $n_F(1, 1)$, and $n_F(2, 1)$ defines all cells and marginal totals for each final table.

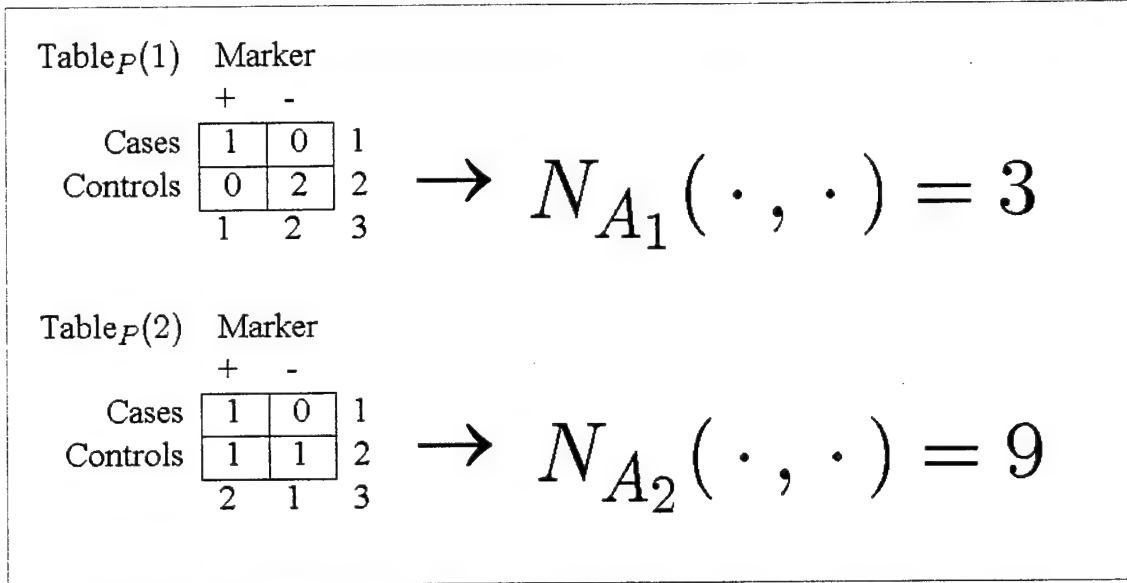


Figure 9. Sample Sizes from Two Different Internal Pilot Study Tables.

1.3.6 Test Statistic

The main purpose for defining a test statistic is to allow us to determine which final tables are considered more extreme and hence determine the appropriate critical values and correct test size based on a specified α level. Without defining a test statistic, we cannot as easily or quickly decide which tables are considered more extreme. The form of the test statistic we are using is Pearson's χ^2 statistic. The benefit of using this form of the test statistic is that we can easily calculate it from each final table and also many statistical software packages will perform this calculation for us. However, since the final sample size is now a random variable because of the sample size re-estimation, it is no longer safe to assume that the final test statistic has an approximate χ^2 distribution. For the i^{th} , j^{th} final table, we define the test statistic $T_{i,j}$ as

$$T_{i,j} = \frac{N_{F_{i,j}}(\cdot, \cdot) [n_{F_{i,j}}(1, 1)n_{F_{i,j}}(1, 2) - n_{F_{i,j}}(2, 1)n_{F_{i,j}}(2, 2)]^2}{N_{F_{i,j}}(1, \cdot)N_{F_{i,j}}(2, \cdot)N_{F_{i,j}}(\cdot, 1)N_{F_{i,j}}(\cdot, 2)}. \quad (25)$$

This statistic will be defined for every final table, $\text{Table}_F(i, j)$.

1.3.7 Critical Value and Power

Define the critical value to be the largest test statistic value such that the probability of rejection is at most α when the null hypothesis is true. The critical value, T_c , is defined to be the value of the test statistic at which the following is true,

$$\Pr\{T \geq T_c\} \leq \alpha. \quad (26)$$

The power can then be defined as simply the probability of rejecting the null hypothesis

$$\text{Power} = \Pr\{\text{rejecting the null}\}. \quad (27)$$

Now, let us define some terms that will be useful in determining the test size inflation. Let α_d be the nominal significance level from using a discrete test statistic distribution based on enumerating every possible 2×2 table. Note, that $\alpha_d \leq \alpha$ by definition. Let α' be the observed significance level based on using an internal pilot study. The test size inflation is determined by the difference between the observed and nominal discrete significance levels, $\alpha' - \alpha_d$.

We have now introduced the problem, provided a motivating scientific example, and defined the notation. In Chapter II, we will derive the exact small sample power and test size for the conditional test, the one where both margins are fixed. In Chapter III, we will derive the exact small sample power and test size for the unconditional test.

The unconditional test assumes that only one margin is fixed, specifically, the number of cases and controls. In Chapter IV, we will provide an example of both the conditional and unconditional approaches to deriving the exact small sample power and test size. Finally, in Chapter V, we will discuss significant results and areas of future research.

Before we present the two approaches, we must discuss the rationale between using an unconditional versus a conditional test. In this thesis, we are not advocating one approach over the other but providing both approaches since we believe the use of each is dictated by the assumptions made about the data. It is then left to the practitioner to determine which approach is more appropriate to use. As stated above, if we have a 2×2 table in which both margins are fixed in advance, the appropriate test would be the conditional one. However, if only one margin is fixed in advance, such as the number of cases and controls, then the unconditional test would be appropriate. Note that many, such as Yates (1984), state that the only rational test, regardless of whether one or both margins are fixed, is the conditional one. We believe that the test used should be dictated by the data and whether one or both margins are truly fixed (or the assumptions made about the margins).

Now, we should also briefly introduce the concept of a nuisance parameter and why it is important in exact inference. In examining 2×2 tables, we have to rid the analysis of any dependence on π , the nuisance parameter in this case. As long as the table probabilities depend on the nuisance parameter, exact inference is not possible. The conditional approach accomplishes this by conditioning on a sufficient statistic for π , the sum of cases and controls with the genetic polymorphism of interest. The unconditional approach frees the analysis of any dependence on π by using the value of π that maximizes the test size over the range of possible π values.

CHAPTER II

THE EXACT CONDITIONAL TEST

In this chapter, we derive the exact small sample distribution of the conditional test statistic for the final table while using an internal pilot study. For a conditional test, the total numbers of cases and controls are fixed. Also, the total number of subjects with the genetic polymorphism is fixed. Therefore, we are conditioning on both margins of the 2×2 table.

With this approach, we enumerate every possible 2×2 table based on the initial sample size estimation (or internal pilot study). Once we have enumerated every internal pilot study 2×2 table, we can re-estimate the final sample size based on the minimally detectable difference of interest and target power. We can then enumerate every possible additional table based on the re-estimated final sample size. This approach is possible because we have a binary response and outcome. Hence, we have a finite number of possible sample sizes and final test statistics, and we can derive the distribution of the final test statistic by determining every possible value of the test statistic and its associated probability. Although this approach becomes quite computationally intensive as the sample size gets large, recent advances in computing allow us to handle rather large sample sizes in a reasonable amount of time.

We first define the sampling scheme and probability structure. Then, we calculate the final table probabilities. Finally, we derive the distribution of the final test statistic.

2.1 Sampling Scheme/Probability Structure

In this section, we define both the sampling scheme and probability structure for the exact conditional test. The sampling scheme and probability structure are identical for the situation with or without an internal pilot study. Therefore, we can define both at the same time. Recall that we have defined the test statistic, T , using the χ^2 test statistic as defined by Everitt (1977)

$$T_{i,j} = \frac{N_{F_{i,j}}(\cdot, \cdot) [n_{F_{i,j}}(1, 1)n_{F_{i,j}}(1, 2) - n_{F_{i,j}}(2, 1)n_{F_{i,j}}(2, 2)]^2}{N_{F_{i,j}}(1, \cdot)N_{F_{i,j}}(2, \cdot)N_{F_{i,j}}(\cdot, 1)N_{F_{i,j}}(\cdot, 2)}. \quad (28)$$

There are just a few ways in which to obtain the same test statistic value. For example, if we observe the 2×2 table shown in Figure 10 Table A, we get a test statistic value of 1.6, which is identical to the test statistic value from the 2×2 table in Table B.

Table A.		Genetic Marker		
		+	-	
Cases	14	2		16
Controls	31	1		32
	45	3		48

Table B.		Genetic Marker		
		+	-	
Cases	16	0		16
Controls	29	3		32
	45	3		48

Figure 10. Example of 2×2 Tables with same Test Statistic Value.

The number of ways in which to obtain the same test statistic value has been reduced by conditioning on the margins. That is, we only need to combine the probabilities from the final tables that lead to the same statistic value if they have the same margins.

2.1.1 Design Issues

In an experiment, we will only observe a single internal pilot study table and final table. However, when we derive the distribution of the final test statistic we must consider all possible tables. The probability structure associated with these tables is dependent on the values of π_1 and π_2 . We can calculate the probability of the table by accounting for the probability of observing a 2×2 table with specified margins $\{N(1, \cdot), N(2, \cdot), N(\cdot, 1), \text{ and } N(\cdot, 2)\}$. The information needed for the margins, however, can be simplified to only $N(\cdot, \cdot)$ and $N(\cdot, 1)$. The reason we can do this is because knowing $N(\cdot, \cdot)$ and $N(\cdot, 1)$ completely defines all the other margins. If we know $N(\cdot, \cdot)$ we can solve for the $N(1, \cdot)$ and $N(2, \cdot)$ since we know the ratio of controls to cases, k . Additionally, knowing $N(\cdot, \cdot)$ and $N(\cdot, 1)$, we can solve for $N(\cdot, 2)$ by subtracting $N(\cdot, 1)$ from $N(\cdot, \cdot)$. Given the margins, we can find the probability of the specific entries in the table, $n(1, 1)$, $n(1, 2)$, $n(2, 1)$, and $n(2, 2)$.

Under our assumptions, some probabilities simplify. The internal pilot study sample size, $N_P(\cdot, \cdot)$, is completely determined by the population values π_1 and π_2 . Therefore, $\Pr\{N_P(\cdot, \cdot)\} = 1$ because we only have one initial estimate of π_1 and π_2 , which will not change. The additional sample size, $N_A(\cdot, \cdot)$, is also completely defined and fixed given the internal pilot study data. Therefore, $\Pr\{N_A(\cdot, \cdot) | \text{Table}_P(i), N_P(\cdot, 1), N_P(\cdot, \cdot)\} = 1$. These two probabilities will help us simplify the probability calculations later in this chapter.

2.1.2 Marginal Distribution

Why do we need to know the probability of the margins? In an experiment, we do not actually account for the marginal probabilities directly, but indirectly through the sampling process. Via the experiment, we observe a specific internal pilot study table and then a specific additional table. However, as stated above, to derive the distribution of the final test statistic, we must consider all possible tables. Therefore, there are many

margins possible for a given sample size. Suppose we have decided that the internal pilot will be half of the initial guess at the sample size. If we have an initial sample size of 12, this would give us an internal pilot study sample size of 6. Table 1 show the set of possible margins given an internal pilot study sample size of 6.

$N_P(\cdot, 1)$	$N_P(\cdot, \cdot)$
0	6
1	6
2	6
3	6
4	6
5	6
6	6

Table 1. All Possible Margins given Internal Pilot Study Sample Size, $N_P(\cdot, \cdot) = 6$.

We must account for the probability of these margins occurring in order to determine the final table probabilities and hence the final test statistic probabilities. Because the total sample size and the ratio of cases to controls are fixed, the problem of determining the marginal probabilities reduces to finding the probability of each first column total, $N_{P_1}(\cdot, 1)$. This total is the sum of the number of cases and controls with the genetic polymorphism of interest.

The number of cases with the genetic polymorphism of interest is binomially distributed with parameters $N_{P_1}(1, \cdot)$ and π_1 . Similarly, the number of controls with the genetic polymorphism of interest is binomially distributed with parameters $N_{P_1}(2, \cdot)$ and π_2 . The total number of subjects with the genetic polymorphism is the sum of two independent binomially distributed random variables. This distribution is known (Ross, 1988, pages 218-219). Thus, for $k \in \{0, 1, \dots, N_{P_1}(\cdot, \cdot)\}$,

$$\begin{aligned}
& \Pr\{N_{P_i}(\cdot, 1) = k | N_{P_i}(\cdot, \cdot)\} \\
&= \sum_{t=0}^{N_{P_i}(1, \cdot)} \binom{N_{P_i}(1, \cdot)}{t} \pi_1^t (1 - \pi_1)^{N_{P_i}(1, \cdot) - t} \times \\
&\quad \binom{N_{P_i}(2, \cdot)}{N_{P_i}(1, \cdot) - t} \pi_2^{N_{P_i}(1, \cdot) - t} (1 - \pi_2)^{N_{P_i}(2, \cdot) - [N_{P_i}(1, \cdot) - t]}.
\end{aligned} \tag{29}$$

Table 2 lists the marginal probabilities for the internal pilot study sample size of 12 under the assumption that $\pi_1 = \pi_2 = 0.1$. We can similarly define the additional probability, $\Pr\{N_{A_{ij}}(\cdot, 1) | N_{A_{ij}}(\cdot, \cdot), \text{Table}_P(i), N_{P_i}(\cdot, 1), N_{P_i}(\cdot, \cdot)\}$, for $k \in \{0, 1, \dots, N_{A_{ij}}(\cdot, \cdot)\}$.

$N_P(\cdot, 1) = m$	$N_P(\cdot, \cdot)$	Probability ($\pi_1 = \pi_2 = 0.1$)
0	12	2.82×10^{-1}
1	12	3.77×10^{-1}
2	12	2.30×10^{-1}
3	12	8.52×10^{-2}
4	12	2.13×10^{-2}
5	12	3.79×10^{-3}
6	12	4.91×10^{-4}
7	12	4.68×10^{-5}
8	12	3.25×10^{-6}
9	12	1.60×10^{-7}
10	12	5.35×10^{-9}
11	12	1.08×10^{-10}
12	12	1.00×10^{-12}

Table 2. Marginal Probabilities given Internal Pilot Study Sample Size, $N_P(\cdot, \cdot) = 12$.

2.1.3 Hypergeometric Distribution

Once we have determined or conditioned on the marginal totals, the probability of each internal pilot study (or additional) 2×2 table can be calculated. Under the null hypothesis, this probability is given by the hypergeometric distribution. The probability of observing a specific internal pilot study 2×2 table given its margins is defined as

$$\Pr\{\text{Table}_P(i) | N_{P_i}(\cdot, 1), N_{P_i}(\cdot, \cdot)\} = \frac{N_{P_i}(1, \cdot)! N_{P_i}(2, \cdot)! N_{P_i}(\cdot, 1)! N_{P_i}(\cdot, 2)!}{N_{P_i}(\cdot, \cdot)! n_{P_i}(1, 1)! n_{P_i}(1, 2)! n_{P_i}(2, 1)! n_{P_i}(2, 2)!} \quad (30)$$

We can similarly define the probability for an additional table given its margins and associated internal pilot study data. The hypergeometric distribution is not appropriate under the alternative hypothesis. The next section describes the appropriate distribution under the alternative hypothesis.

2.1.4 Adjusted Product Binomial

Under the alternative hypothesis, the cases and controls are assumed to have different proportions with the genetic polymorphism of interest. In order to properly account for these different proportions we must use a method that uses the proportions in the probability calculation. One method would be to use a form of the product binomial, as described in the Proc-StatXact User Manual (Cytel, 1999). First, we must define the notation necessary for calculating this probability and the conditions for which it is appropriate. For the internal pilot study, the total sample size is already fixed. Therefore, we need only account for the first column total, $N_{P_i}(\cdot, 1)$. It is necessary to group those 2×2 tables with the same marginals together, so let Γ_m denote all enumerated internal pilot study tables, $\text{Table}_P(i)$, which have a total m polymorphism. That is

$$\Gamma_m = \{i : N_{P_i}(\cdot, 1) = m\}. \quad (31)$$

We can now define the probability for each internal pilot study table accounting for the probability of the cases, where $t = 1$, and controls, where $t = 2$, explicitly as

$$\Pr\{\text{Table}_P(i) | N_{P_i}(\cdot, 1) N_{P_i}(\cdot, \cdot)\} = \frac{\prod_{t=1}^2 \left(\frac{N_{P_i}(t, \cdot)}{n_{P_i}(t, 1)} \right) \pi_t^{[n_{P_i}(t, 1)]} (1 - \pi_t)^{[N_{P_i}(t, \cdot) - n_{P_i}(t, 1)]}}{\sum_{N_{P_i}(\cdot, 1) \in \Gamma_m} \left[\prod_{t=1}^2 \left(\frac{N_{P_i}(t, \cdot)}{n_{P_i}(t, 1)} \right) \pi_t^{[n_{P_i}(t, 1)]} (1 - \pi_t)^{[N_{P_i}(t, \cdot) - n_{P_i}(t, 1)]} \right]}. \quad (32)$$

We can similarly define the probability for an additional table given its total sample size, $N_{A_{i,j}}(\cdot, \cdot)$, its first column total, $N_{A_{i,j}}(\cdot, 1)$, and the internal pilot study data as $\Pr\{\text{Table}_A(i, j) | N_{A_{i,j}}(\cdot, 1), N_{A_{i,j}}(\cdot, \cdot), \text{Table}_P(i)\}$. Table 3 provides a simple example using a portion of data from an internal pilot study of size 12 under the assumption that $\pi_1 = \pi_2 = 0.1$.

2.2 Exact Small Sample Power for the Conditional Test

Now, we have defined all the pieces necessary to calculate the probability of observing a specific final table. Now, we define the final table probabilities. We can determine the probability of a specific test statistic value by summing the probability of all final tables that lead to the same test statistic value under the same conditions and from the same internal pilot study table. This in turn allows us to determine the distribution of the final test statistic by assigning a probability to each unique value of the test statistic. Finally, we show how to calculate the power while using an internal pilot study and this approach.

$N_P(\cdot, 1)$	$N_P(\cdot, \cdot)$	$n_P(1, 1)$	$n_P(1, 2)$	$n_P(2, 1)$	$n_P(2, 2)$	Adjusted Probability
0	12	0	4	0	8	1.000
1	12	0	4	1	7	0.667
1	12	1	3	0	8	0.333
2	12	0	4	2	6	0.424
2	12	1	3	1	7	0.485
2	12	2	2	0	8	0.091
3	12	0	4	3	5	0.255
3	12	1	3	2	6	0.509
3	12	2	2	1	7	0.218
3	12	3	1	0	8	0.018
...
10	12	2	2	8	0	0.091
10	12	3	1	7	1	0.485
10	12	4	0	6	2	0.424
11	12	3	1	8	0	0.333
11	12	4	0	7	1	0.667
12	12	4	0	8	0	1.000

Table 3. Example of Adjusted Product Binomial Probabilities for $\pi_1 = \pi_2 = 0.1$.

2.2.1 Final Table Probability

The multiplication rule (Mood, Graybill, and Boes: 1974, page 37) provides a general equation for finding the probability of each final table. This probability is required to determine the distribution of the final test statistic. We can write the final table probability, $\Pr\{\text{Table}_F(i, j)\}$, as the joint probability of $\text{Table}_P(i)$ and $\text{Table}_A(i, j)$, such that

$$\Pr\{\text{Table}_F(i, j)\} = \Pr\{\text{Table}_P(i), \text{Table}_A(i, j)\}. \quad (33)$$

The probability of observing a specific final table, $\text{Table}_F(i, j)$, is given by

$$\Pr\{\text{Table}_F(i, j)\} = \Pr\{\text{Table}_P(i)\}\Pr\{\text{Table}_A(i, j)|\text{Table}_P(i)\}. \quad (34)$$

Now that we have defined the probability of observing a specific final table, we can then define the test statistic distribution and power. First, we will discuss the test statistic distribution.

2.2.2 Test Statistic Distribution

Recall from § 1.3.6, the definition of the test statistic is

$$T_{i,j} = \frac{N_{F_{i,j}}(\cdot, \cdot) [n_{F_{i,j}}(1, 1)n_{F_{i,j}}(1, 2) - n_{F_{i,j}}(2, 1)n_{F_{i,j}}(2, 2)]^2}{N_{F_{i,j}}(1, \cdot)N_{F_{i,j}}(2, \cdot)N_{F_{i,j}}(\cdot, 1)N_{F_{i,j}}(\cdot, 2)}. \quad (35)$$

It is possible for some tables with the same margins to have the same test statistic value. There are a finite number of final tables. There are also a finite number of unique test statistic values, which is less than or equal to the number of final tables. Let us write the unique set of test statistic values for each set of margins that came from the same internal pilot study table i as $T_{i,u}$ where $u \in \{1, 2, \dots, U_i\}$, where U_i is the number of unique test statistic values that are possible from the i^{th} internal pilot study. Using this notation allows us to have a different set of unique test statistic values for each internal pilot study table which is necessary due to the conditioning. The probability of observing a specific test statistic value, $\{T_{i,1}, T_{i,2}, \dots, T_{i,u}\}$, is the sum of the probabilities of all the final tables that lead to that same test statistic value and have the same margins,

$$\Pr\{T_{i,u}\} = \sum_{j=1}^m \Pr\{\text{Final Table } (i, j)\} I\{\text{Final Table } (i, j) \in T_{i,u}\}, \quad (36)$$

where the $\Pr\{\text{Final Table } (i, j)\}$ is defined in Equation 34 and

$I\{\text{Final Table}(i, j) \in T_{i,u}\}$ is an indicator variable for those final tables that lead to the same unique test statistic value.

An example, starting with an initial total sample size of 24, is provided below in Table 4. As one can see from the table, we have multiple test statistic values that are equal and others that are unique. We sum the probabilities of these identical test statistic values which gives us the probability of observing the unique test statistic value. Now, we have a unique set of test statistic values and their associated probabilities. This is the distribution of the final test statistic.

Test Statistic Value	$\Pr\{\text{Final Table } (i, j)\}$	Sum of $\Pr\{\text{Final Table } (i, j)\}$
6.86	4.38×10^{-4}	4.38×10^{-4}
9.6	1.22×10^{-5}	2.07×10^{-4}
	1.95×10^{-4}	
3.75	3.89×10^{-4}	7.78×10^{-4}
	3.89×10^{-4}	
12.63	5.40×10^{-5}	3.78×10^{-5}
	3.24×10^{-5}	
	1.08×10^{-5}	
6.19	1.73×10^{-4}	3.67×10^{-4}
	1.08×10^{-5}	
	1.73×10^{-4}	
3.16	7.56×10^{-5}	7.57×10^{-5}
2.02	3.46×10^{-4}	4.97×10^{-4}
	1.51×10^{-4}	
16	9.01×10^{-7}	3.30×10^{-6}
	2.40×10^{-6}	
	4.80×10^{-6}	
9	2.88×10^{-5}	6.72×10^{-5}
	4.80×10^{-6}	
	2.88×10^{-4}	

Table 4. Example of Multiple Test Statistic Values.

Now that we have defined the probability of a specific final table and the distribution of the final test statistic, we can now define the critical value and power. We will define the power first since it is needed to define the critical value.

2.2.3 Power and the Critical Value

Recall the definition of power is simply the probability of rejecting the null hypothesis. Power always depends on the distribution of the test statistic because the rejection region and critical value are defined in term of this distribution. In order to determine the rejection region, we need to know the critical value as defined in § 1.3.7. However, in order to determine the critical value, we must first develop a methodology for calculating the power. The reason for this is that in order to determine the critical value in the discrete case, we must calculate the power under the null hypothesis. Once we know the critical value, we can then evaluate the power elsewhere. The difference between the continuous and discrete case is that for the continuous case there is a closed form solution to finding the critical value; this is not true for the discrete small sample exact case. Note that the power under the null hypothesis is the test size; a fact we will make use of later to determine the test size inflation.

For all values of the test statistic, we need to know whether we reject the null hypothesis or not. That is, is $T_{i,u} \in \text{Rejection Region}$. For any given test statistic value, there are only one of two possibilities, either the test statistic falls in the rejection region or it does not. We can define the following indicator function as

$$I\{T_{i,u} \in \text{Rejection Region}\} = \begin{cases} 0 & \text{if } T_{i,u} \notin \text{Rejection Region} \\ 1 & \text{if } T_{i,u} \in \text{Rejection Region} \end{cases} \quad (37)$$

The test statistic values, $T_{i,u}$, are grouped not only by what we are conditioning on, but also by which internal pilot study table they came from, that is, for each set of margins within those final tables that came from the same internal pilot study table i . Therefore,

as shown in the equation above, each internal pilot study table i will have its own set of unique test statistic values, U_i . This gives us the appropriate probability in the rejection region for each internal pilot study i given the margins of the final tables. Recall from § 1.3.5, B is the total number of possible internal pilot study tables. We are now ready to define Theorem 1.

Theorem 1:

$$\text{Power} = \sum_{i=1}^B \sum_{u=1}^{U_i} \Pr\{T_{i,u}\} I\{T_{i,u} \in \text{Rejection Region}\}. \quad (38)$$

Proof: Power is defined as the probability of rejecting the null hypothesis. We have defined a situation in which the probabilities of the test statistic values are summed over the set that are rejected or fall in the rejection region. Therefore, Theorem 1 follows. \square

Equation 38 is a general formula for determining the power and will be used to help define our critical value in the next subsection. The critical value is needed to determine the power under situations other than under the null. Recall from § 1.3.7 that the critical value, T_c , is defined as

$$\Pr\{T \geq T_c\} \leq \alpha. \quad (39)$$

For instance, if we set the nominal significance level, α , at 0.05, then we would want the probability of the rejection region under the null to be no greater than 0.05.

Therefore, we can solve for the critical value, T_c , as the largest test statistic value such that the probability that we reject the null hypothesis is at most α when it is in fact true.

Recall, that we are conditioning on the margins and we will, therefore, have a critical value for each set of margins.

Now, as an example, we will use a nominal significance level, α , of 0.05 and total initial sample sizes of 24, 48, and 96. Note that in order to find the critical value, it will be necessary to calculate the discrete significance level, α_d , which is a byproduct of finding the critical value. Additionally, to show that the test size is inflated, we will need to find the critical value and discrete significance level, α_d , without using an internal pilot study as a comparison. The results for a sample size of 24 are shown in Figures 11 and 12. In Figures 11 and 12, the x -axis is the condition on which the critical values and discrete significance levels or test sizes are based. In this particular example, the final sample size is fixed so the only condition is the first column total, $N(\cdot, 1)$, or number of cases and controls with the genetic polymorphism of interest. Note that not all of the conditions are listed here. The reason for this is that for some conditions, no critical value or discrete test size exists which would give us a probability below 0.05. Also note that for the least extreme marginal totals, $N(\cdot, 1)$, the critical values are the smallest and, correspondingly, the test sizes are closest to the nominal significance level. Now that we have calculated the critical values and discrete test sizes without using an internal pilot study, we can use these as a comparison. The discrete test sizes are used to calculate the test size inflation shown in the next subsection.

Example power plots using initial sample sizes of 24, 48, and 96 are provided below which examine the power at $\pi_2 = 0.1, 0.3, 0.5$, see Figures 13-15. These plots were created using the critical values based on a study without an internal pilot, e.g. Figure 11 for a sample size of 24. Once we have defined the possible critical values without using an internal pilot study, we can then use these values to actually determine the power in the instance when we are using an internal pilot study. In order to

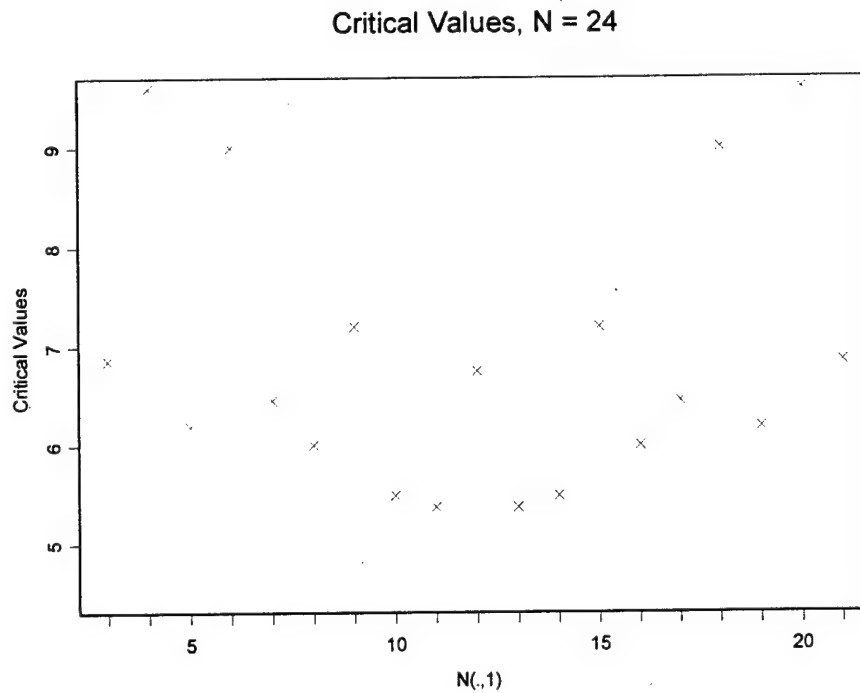


Figure 11. Critical Values Without an Internal Pilot Study, $N = 24$.

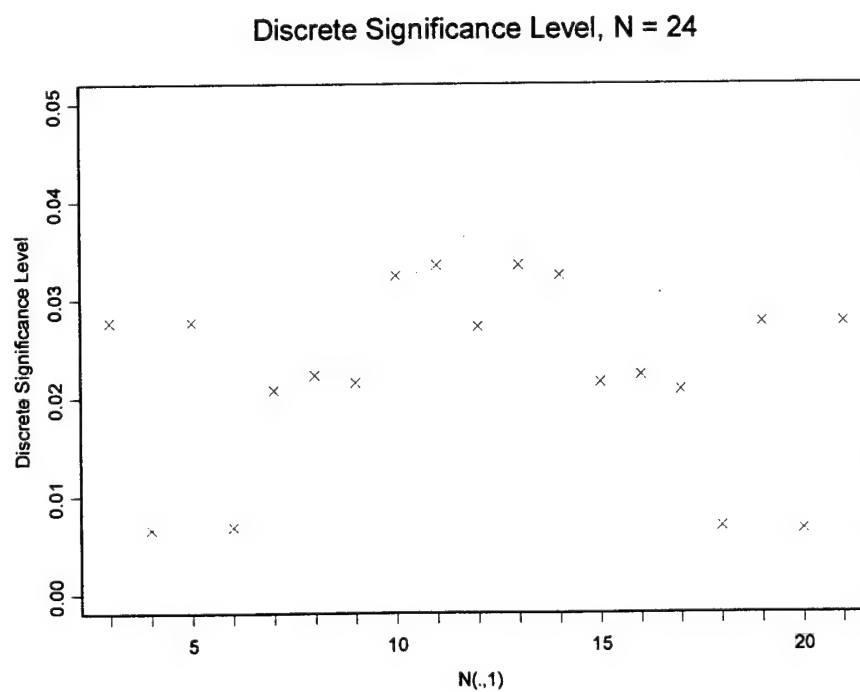
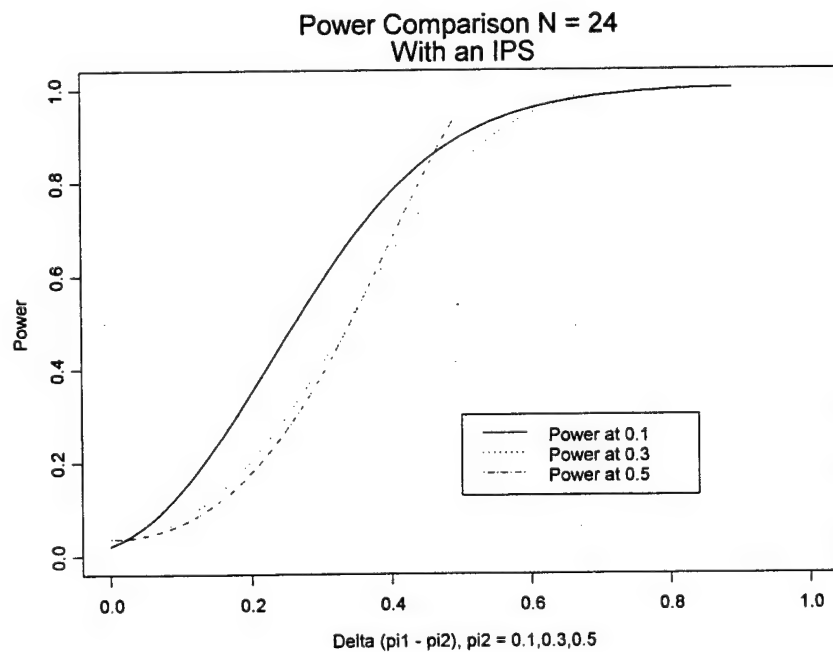


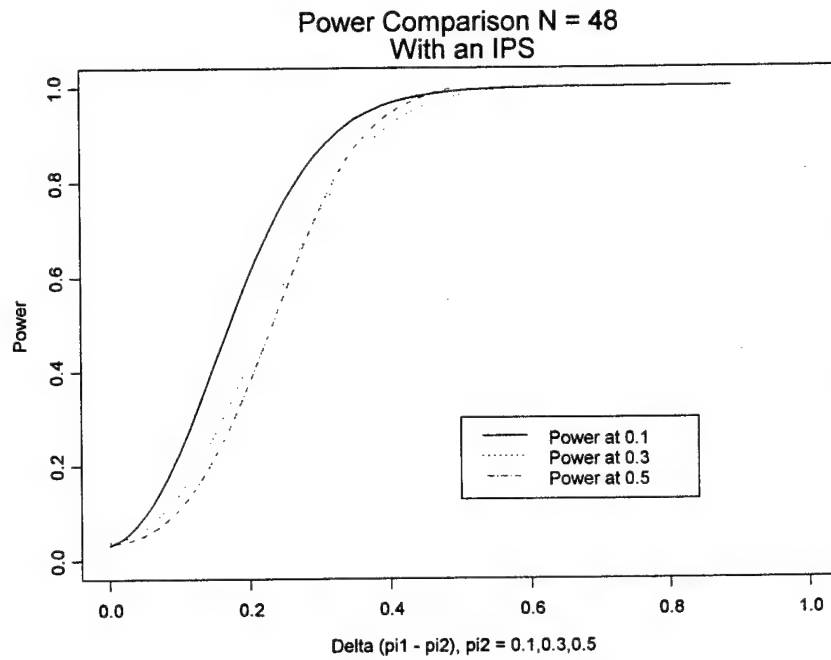
Figure 12. Discrete Significance Levels Without an Internal Pilot Study, $N = 24$.

determine the power, we first need to enumerate every possible table based on our internal pilot and additional sample sizes. Combining the internal pilot and additional sample sizes to get our final sample sizes, we can then decide whether to reject the null hypothesis or not based on our previously calculated critical values (based on without an internal pilot study) and on our conditions, $N(\cdot, 1)$ and $N(\cdot, \cdot)$. Therefore, we can determine whether to reject each final table by comparing it with the associated critical value. The power can then be calculated by summing the probabilities of those tables in which we reject the null hypothesis. Now, we can show that the test size is actually inflated when using an internal pilot study to re-estimate the final sample size using a conditional approach.



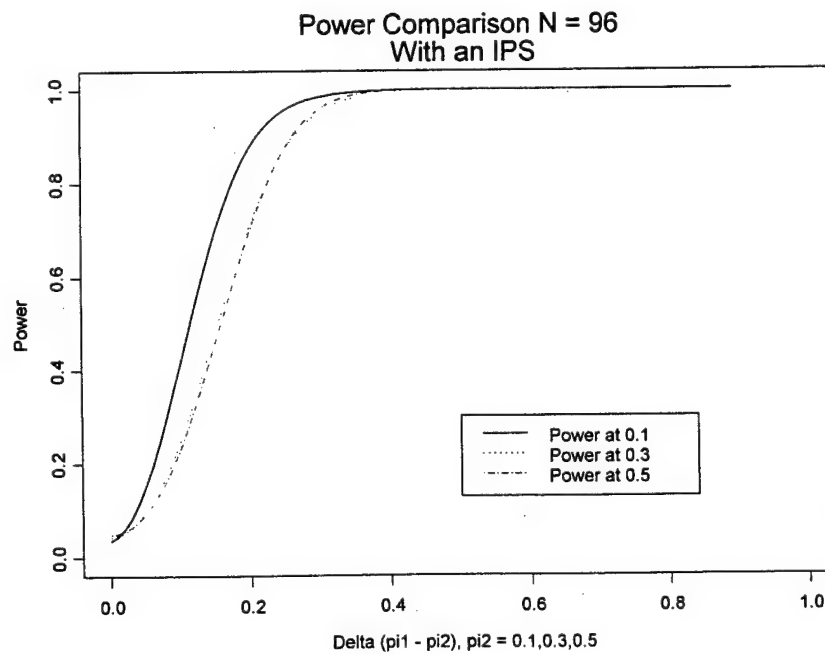
Note: We are limiting the δ ($\pi_1 - \pi_2$) to positive values.

Figure 13. Example Power Plots for Initial Sample Size of 24.



Note: We are limiting the $\delta (\pi_1 - \pi_2)$ to positive values.

Figure 14. Example Power Plots for Initial Sample Size of 48.



Note: We are limiting the $\delta (\pi_1 - \pi_2)$ to positive values.

Figure 15. Example Power Plots for Initial Sample Size of 96.

2.2.4 Test Size Inflation

As stated previously, the test size inflation can be determined by subtracting the discrete significance level, α_d , from the observed test size, α' . The results are provided in Table 5. Note that using this method, we do not have a single critical value against which to compare our enumerated data since there is a different critical value under each condition for each internal pilot study i . Using this conditional approach, we have a possibly different critical value for each set of margins for each internal pilot study i . Only values below 0.5 are shown since those above are the mirror image. In Table 5, notice that the test size inflation is largest at $N = 96$, but there is significant overlap between the 3 sample sizes compared. The general pattern to note for this table is that for all values examined, there is an inflation in test size, as shown in Table 5 and Figure 16. Note that we have only shown an inflation in test size empirically. The important message from both Table 5 and Figure 16 is that the test size inflation is above zero and the observed test size is below 0.05. A more complete table of inflation values is provided in Appendix B.

	0.1	0.2	0.3	0.4	0.5
24	13.80	15.50	11.20	8.77	8.72
N 48	10.40	9.72	12.80	14.80	6.30
96	6.81	8.10	7.96	8.21	19.20

Table 5. Test Size Inflation ($\times 10^{-3}$) with an Internal Pilot Study
Sample Size = 24, 48, 96. Test size inflation calculated by subtracting the
significance level without an internal pilot study, α_d , and the significance level
with an internal pilot study, α' . No α' exceeds 0.05.

The most important finding in this section is that for small samples with $\alpha = 0.05$, we can get the benefit of using an internal pilot study without paying any penalty! The reason for this is that in small samples, the discreteness pushes the discrete significance level so far below the nominal significance level of 0.05 that even

with the inflation from using an internal pilot study, we are still below the nominal significance level, as an example see Figure 17. This is not as obvious in Table 5 and Figure 16 since we have subtracted the discrete significance level without an internal pilot from the discrete significance level with an internal pilot study. Therefore, in small samples, no adjustment is necessary when using an internal pilot study. An important item of interest would be the sample size point at which this breaks? At what sample size do we start paying a penalty for using an internal pilot study? For example, if we assume the maximum sample size to be four (4) times that initially estimated instead of just twice, we can actually calculate where the test size inflation is above our desired significance level, e.g. 0.05.

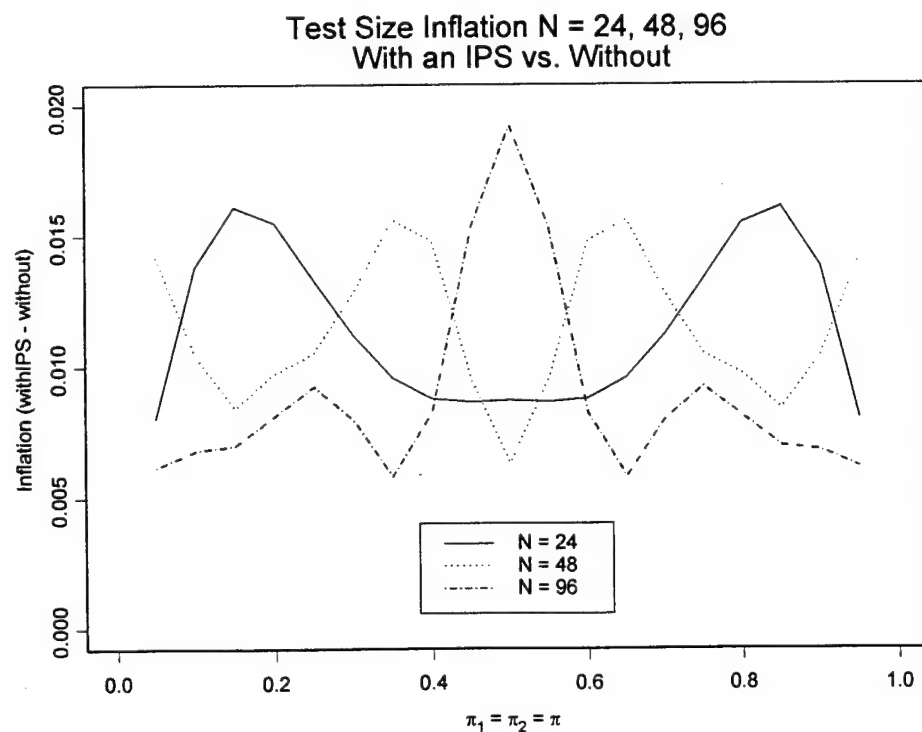


Figure 16. Test Size Inflation.

2.2.5 Adjusting for Test Size Inflation

In this section, we discuss what to do when the test size is inflated from using an internal pilot study. As shown in a § 2.2.4, no adjustment is necessary for small sample sizes, however, with larger sample sizes it is necessary because the inflation is large enough to increase the test size above the nominal significance level. This is a result of the product binomial distribution becoming less and less discrete as the sample size increases which allows the discrete significance level or test size, α_d , to get closer and closer to the nominal significance level, α . Since the discrete significance level is getting closer to the nominal, any inflation pushes the test size above the nominal significance level. There does not appear to be any direct correlation between the sample size and the amount of inflation as shown in Figure 16.

As previously mentioned, we cannot simply correct for the inflation by changing the critical value because there are many critical values to consider, a result of the conditioning arguments used to derive the distribution of the test statistic. However, the manner in which we determine the critical values provides us with a means to compensate for the inflation by adjusting the nominal significance level, α , before applying the test. Simply stated, the nominal significance level is reduced to the largest value such that the resulting discrete significance level or test size, α_d , is below the nominal one. If in fact the discrete significance level is below the nominal, there is no need to adjust for inflation. Note that even with the adjustment, there is still an inflation in test size when using an internal pilot study; however, the adjustment brings the test size below the nominal significance level.

The adjustment procedure involves lowering the nominal significance level during the critical value and discrete significance level calculations only, and not also in the initial sample size calculation or recalculation process. This means that the probability that $T \geq T_c$ is less than or equal to the adjusted nominal significance level.

The same ratio of controls to cases is maintained. We will limit the reductions in the nominal significance level to 0.001 increments.

As an example, Figure 17 shows a test size inflation at an initial sample size of 129 that is above the nominal significance level of 0.05. In order to adjust the test size so that it is below the nominal, we must recalculate the critical values and discrete significance levels by lowering the nominal significance level. Table 6 shows the resulting discrete significance levels or test sizes for two adjusted nominal significance values.

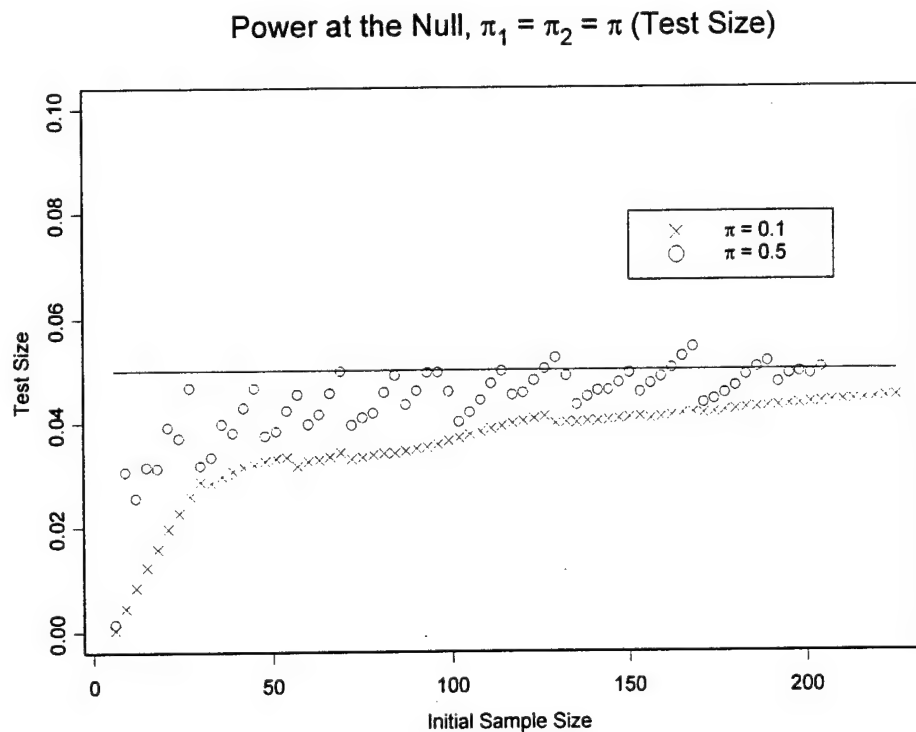


Figure 17. Where Test Size Adjustment is Required.
 Notice that even at a sample size of over 200 using a $\pi = 0.1$ there is still no need to adjust the test size. This is not the case for $\pi = 0.5$.

Nominal Significance Level, α	Discrete Significance Level, α_d
0.05	0.0524
0.049	0.0523
0.048	0.0494

Table 6. Example of Adjustment Procedure at the Null, $\pi_1 = \pi_2 = 0.5$. Initial Sample Size is 129. The discrete nominal significance level is 0.0396. Even with adjustment, there is still inflation.

Keep in mind that what we have done is empirical. At this point there does not appear to be a simpler or approximate manner in which to compensate for the inflation above the nominal significance level. In many instances when using a discrete distribution, it is possible to approximate a solution. For example, when analyzing a 2×2 table, we can use Fisher's exact or a normal approximation given the expected cell counts are large enough. But, what should we do when the expected cell counts are not large enough? In this instance, it is not appropriate to use a normal approximation. The same is true for the conditional enumeration approach proposed in this chapter. However, when using an enumeration type of approach and the conditional arguments, there are always subsets of the data that do not meet the assumptions required to approximate the distribution with a continuous one.

2.2.6 Sample Size Issues

When we first discussed the study design, we suggested using the sample size equation proposed by Rosner (1995, page 384) to estimate the sample size. Using the ideas developed in Chapter II, we can now suggest a more appropriate technique. We propose a new and better approach to sample size estimation in small samples. The new approach is an iterative one that will be more accurate and more appropriate. The new approach is exact and therefore more consistent with the exact tests and exact power calculations. However, this new approach can be a big computing problem

because of its iterative nature. However, prior to submitting any papers for publication, we will implement this alternative approach to the sample size calculations.

At all times, we should keep in mind that the maximum allowable sample size will depend on the investigators; but might be further limited by money, finite population sizes, ethical reasons, or time. The investigators should account for these when determining the maximum allowable sample size. If we are already at this maximum and the sample size is too small, we should stop there. We do not want to limit our new iterative sample size approach at this point by setting the sample size too small. Next, we will briefly outline the proposed alternative approach to the sample size calculations which includes a binary search algorithm (Harel, 1992).

The proposed alternative begins with the sample size equation from Rosner for the difference in two binomial proportions. Once we have an initial estimate, we perform a power analysis as though we are not considering an internal pilot study, Figure 18, point $(N_1, Power_1)$. There are three possible outcomes after determining the initial sample size and estimated power. The sample size may be too big, and hence the power is too big. Alternatively, the sample size maybe too small, and hence the power is too small. The sample size is okay, and the power is within the pre-specified tolerance.

As an example, we have assumed that the first iteration produced an initial sample size that is too big $(N_1, Power_1)$. Therefore, the next iteration takes half the distance between the initial sample size and zero. For this iteration, the expected power is too small $(N_2, Power_2)$. Therefore, the next iteration will half the distance between the second sample size and the initial sample size $(N_3, Power_3)$, since the initial sample size in this example was too large. Now, assuming $(N_3, Power_3)$ is too big again, we would need at least another iteration, $(N_4, Power_4)$. This is what is meant by a binary search algorithm (Harel, 1992). The algorithm continues until we obtain an

expected power within a specified tolerance of the goal power. See Figure 18 for a graphical representation of the above example using the binary search algorithm.

Since we are using an internal pilot study, we will need to perform two sample size calculations just as with the simplified approach. However, as mentioned above, this alternative approach will be iterative and therefore more difficult and time consuming. The proposed alternative begins with the sample size equation for comparing two binomial proportions. Once we have an estimate of the initial sample size, we can then enumerate every possible table and calculate the power as done in § 2.2.4. Once we have a power estimate, we can then determine if our sample size is large enough or not. There are three possibilities. The power is greater than desired (sample size is too big). The power is smaller than desired (sample size is too small). Finally, the power is within the pre-specified tolerance. If the sample size is too large, we know that we have captured the desired sample size between what was calculated and zero. If the sample size is too small, we know that the desired sample size is greater than what was calculated. In these two cases, we are left with essentially the same problem, how much to add or subtract. It should also be noted that there are many ways in which to determine the amount to add or subtract, however, in order to keep the solution as simple as possible, we think the simplest approach will be to just use the binary search algorithm described in the paragraph above. Once the new sample size is determined, we can follow the same pattern as before: 1) enumerate every possible table, 2) calculate the power, and 3) compare against the target power. This will continue until we have determined a sample size that achieves the desired power. Keep in mind, this is only for the initial sample size. We still need to re-estimate the final sample size after collecting the internal pilot study data.

Sample Size Calculations - Power too Small

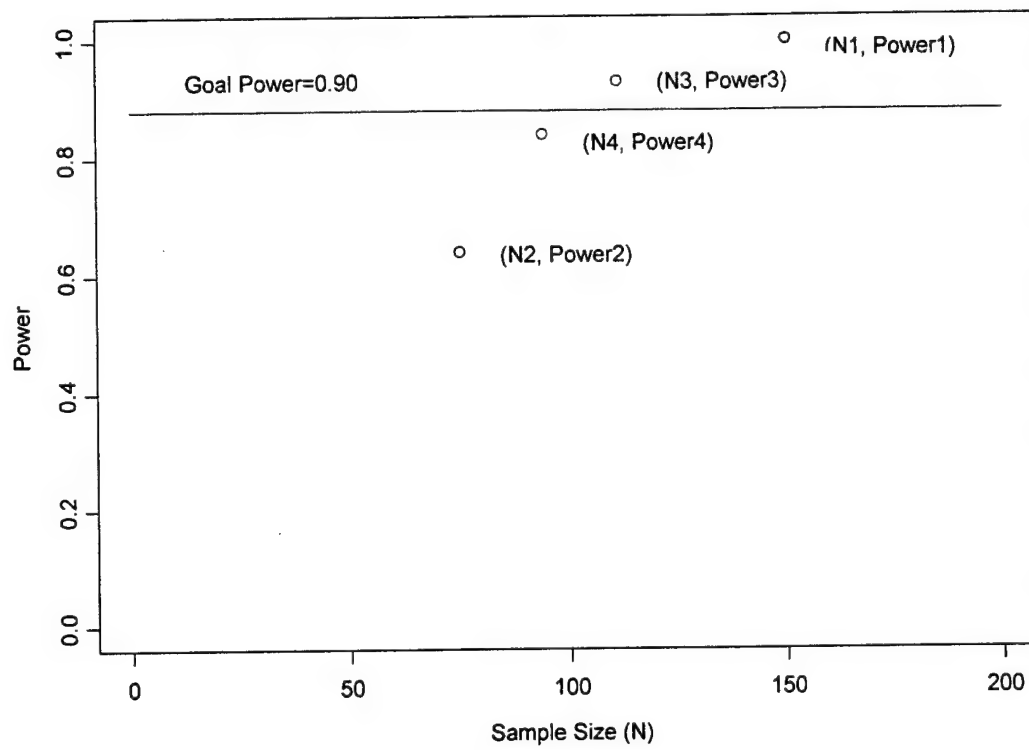


Figure 18. Iterative Sample Size Approach for Initial Sample Size.

Once the internal pilot study data has been collected, we will need to re-estimate the final sample size based on what we observe from the internal pilot study data. We will follow the same process as outlined above. Based on our internal pilot study data, we will update our estimates of the population parameters, π_1 and π_2 . With these new parameter estimates, we can use the sample size equation for comparing two binomial proportions as the starting point. Once again, we will enumerate every possible additional table and determine the power as outlined in § 2.2.4. As before, we will use the binary search algorithm to determine the next sample size to try. This process will continue until we converge on the appropriate sample size that gives us the desired power.

2.3 Summary and Conclusions

We have now provided the analyst and practitioner a means of analyzing a case-control study with an internal pilot study using a conditional approach. This approach is consistent with the literature in terms of using a conditional approach regardless of the type of data.

Our approach is consistent with the literature in terms of using a conditional test. By using an internal pilot study, however, we have introduced the possibility of test size inflation. However, the most significant result from this section is that with small sample sizes, we have the benefit of using an internal pilot study without the drawback of test size inflation above the nominal significance level. However, if there is test size inflation above the nominal, we have provided a method to adjust the inflation to ensure the test size stays below the nominal. This is an exact method which also allows us to produce exact p -values.

CHAPTER III

THE EXACT UNCONDITIONAL TEST

In this section, we derive the small sample exact distribution of the final sample size and hence the final test statistic while using an internal pilot study and the unconditional test. In the unconditional test we condition on only one set of margins of the 2×2 table, unlike in the conditional test in which we condition on both sets of margins. Specifically, we will condition only on the number of cases and controls.

The approach, once again, is to enumerate every possible table based on the initial sample size estimation (or internal pilot study) and the additional data. For each enumerated internal pilot study table, we re-estimate the final sample size required to maintain a power of 90% and a Type I error rate of 0.05. From this, we determine the additional sample size. Now, we can enumerate every possible additional table. Thus, we can derive the distribution of the final sample size and hence the final test statistic by determining every possible value of the test statistic and its associated probability just as we did in Chapter II. However, the main difference is in the probability calculation for each test statistic value. Although the test statistic values are identical whether we use the conditional or unconditional approach, the probability structures differ. The probability of observing a specific test statistic value will change based on which approach is used, conditional or unconditional.

Recall from § 1.2.2, Suissa and Shuster (1985) derived an exact approach to unconditional tests. They used a Z statistic based on their 2×2 contingency tables assuming two independent binomial samples of equal size. This allowed them to use the product binomial distribution to determine their Z statistic probabilities. We plan to use an approach almost identical to that of Suissa and Shuster; that is, our data are in

the form of a 2×2 contingency table and we also assume two independent binomial samples for each 2×2 table. We also use a product binomial distribution to determine the test statistic probabilities. We do not, however, assume equal sizes.

In their approach, Suissa and Shuster would have set the proportions of cases and controls with the genetic polymorphism to be equal, e.g. $\pi_1 = \pi_2 = \pi$. Under this null, they calculate the observed test size, α_d , based on using the product binomial distribution. They then find the value of π that maximizes the observed test size, α_d , using a grid search technique for the value of π between 0.001 and 0.999. The maximum value of α_d over all of the π 's is then considered the test size for this unconditional test, $\alpha_{d_{max}}$. Treating π as a nuisance parameter and removing it from the problem, as just described, leads to unconditional inference. The reason this is true is because we are no longer conditioning on a specified value of π . Instead, we maximize α or test size over all values of π . Unconditional inference requires sample sizes smaller than the exact conditional counterpart since the hypothesis test is less conservative.

Before we begin with the derivation of the final test statistic, we must, once again, define the sampling scheme and probability structure. Then, we can calculate the final table probabilities and hence the distribution of the final test statistic.

3.1 Sampling Scheme/Probability Structure

This section will define both the sampling scheme and probability structure. The definition of the test statistic has not changed from § 1.3.6.

In the unconditional approach, as in the conditional one, we assume that the cases and controls are independent. Also, we are only conditioning on the total number of cases and controls, $N(1, \cdot)$ and $N(2, \cdot)$, respectively. Based on this information, we know that the numbers of cases and controls with the genetic polymorphism of

interest are distributed as follows: $n_P(1, 1) \sim \text{Binomial}(N(1, \cdot), \pi_1)$ and similarly $n_P(2, 1) \sim \text{Binomial}(N(2, \cdot), \pi_2)$.

The probability of observing a specific internal pilot study 2×2 table given its total sample size, $N_P(\cdot, \cdot)$, and the values of π_1 and π_2 is defined as

$$\begin{aligned} \Pr\{\text{Table}_P(i) | N_{P_i}(\cdot, \cdot), \pi_1, \pi_2\} = \\ \left(\frac{N_{P_i}(1, \cdot)}{n_{P_i}(1, 1)} \right) \pi_1^{[n_{P_i}(1, 1)]} (1 - \pi_1)^{[N_{P_i}(1, \cdot) - n_{P_i}(1, 1)]} \times \\ \left(\frac{N_{P_i}(2, \cdot)}{n_{P_i}(2, 1)} \right) \pi_2^{[n_{P_i}(2, 1)]} (1 - \pi_2)^{[N_{P_i}(2, \cdot) - n_{P_i}(2, 1)]}. \end{aligned} \quad (40)$$

Equation 40 defines the final table probabilities under both the null and alternative hypotheses.

We can similarly define the probability for an additional table given its total sample size, $N_{A_{i,j}}(\cdot, \cdot)$, and internal pilot study data as

$\Pr\{\text{Table}_A(i, j) | N_{A_i}(\cdot, \cdot), \text{Table}_P(i), N_{P_i}(\cdot, \cdot), \pi_1, \pi_2\}$. Recall from § 2.2.1, we have a general means of calculating the final table probabilities. Finally the probability of observing a final table is

$$\begin{aligned} \Pr\{\text{Table}_F(i, j) | \pi_1, \pi_2\} = \\ \Pr\{\text{Table}_P(i) | N_{P_i}(\cdot, \cdot), \pi_1, \pi_2\} \times \\ \Pr\{\text{Table}_A(i, j) | N_{A_i}(\cdot, \cdot), \text{Table}_P(i), N_{P_i}(\cdot, \cdot), \pi_1, \pi_2\}. \end{aligned} \quad (41)$$

Note that the probability of each final table, $\Pr\{\text{Table}_F(i, j) | \pi_1, \pi_2\}$, is dependent on the values of π_1 and π_2 .

3.2 Exact Small Sample Power for the Unconditional Test

The final table probabilities, test statistic distribution, power and critical value are defined identically as in § 2.2 and therefore will not be repeated in this section. The only difference is that the table probabilities are calculated using a product binomial distribution as described in the section above.

3.2.1 Critical Values and the Nuisance Parameter

In this section, we describe the exact unconditional test and how to determine the value of the nuisance parameter for obtaining critical values and test size. As in Chapter II, we use half of the initial sample size for the internal pilot study. For different initial sample sizes, it is possible for the value of the nuisance parameter, π , to be different. Therefore, we must determine the appropriate value of the nuisance parameter in each instance. This is done by enumerating every possible internal pilot study table and additional table under the null assuming different values of the nuisance parameter, $\pi_1 = \pi_2 = \pi$. Then, just as done by Suissa and Shuster (1985), we find the value of the nuisance parameter that maximizes the observed test size, α' , under the null hypothesis using a grid search technique for values of π from 0.001 to 0.999.

First, however, we must define the critical values and how they relate to the values of π_1 and π_2 . Recall from Chapter II that the critical value is defined based on a final sample size. This is also true for the unconditional test; however, there is also another dimension to the critical value. The critical value also depends on the value of the nuisance parameter, π . The critical value is defined to be the largest test statistic value such that the probability of rejection is at most α when the null hypothesis is true for a given $N_F(\cdot, \cdot)$ and π . Therefore, the critical value in the unconditional test is defined as T_{c_π} since the probability associated with the critical values depends on the value of the nuisance parameter, π . The critical value is the largest test statistic value such that the following is true,

$$\Pr\{T \geq T_{c_\pi}\} \leq \alpha. \quad (42)$$

For a given final table, we need to know whether we reject the null hypothesis or not. That is, is $\text{Table}_F(i, j) \in \text{Rejection Region}$ for a given π ? Recall that for each internal pilot study table, there are a number of additional/final tables possible. For any given final table, there are only one of two possibilities, either the table falls in the rejection region or it does not. This now depends on the value of π . We can define the following indicator function as

$$\begin{aligned} I\{\text{Table}_F(i, j) \in \text{Rejection Region}|\pi\} \\ = \begin{cases} 0 & \text{if } \text{Table}_F(i, j) \notin \text{Rejection Region}|\pi \\ 1 & \text{if } \text{Table}_F(i, j) \in \text{Rejection Region}|\pi \end{cases} \end{aligned} \quad (43)$$

Recall that B is the total number of possible internal pilot study tables. Let $i \in \{1, 2, \dots, B\}$ index each possible internal pilot study table. Recall that M_i is the number of additional/final tables that result from the i^{th} internal pilot study table. Let $j \in \{1, 2, \dots, M_i\}$ index each possible additional/final table. We can now define the observed test size as a function of the nuisance parameter, $\pi_1 = \pi_2 = \pi$, as

$$\alpha'_\pi = \sum_{i=1}^B \sum_{j=1}^{M_i} \Pr\{\text{Table}_F(i, j)|\pi\} I\{\text{Table}_F(i, j) \in \text{Rejection Region}|\pi\} \quad (44)$$

where $\Pr\{\text{Table}_F(i, j)|\pi\}$ and $I\{\text{Table}_F(i, j) \in \text{Rejection Region}|\pi\}$ are defined in Equations 41 and 43, respectively.

We are now prepared to define the maximum value of α'_π . First, let us define an interval I such that $I = [0.001, 0.999]$ by 0.001. The maximum value of α'_π is then defined as

$$\alpha'_{max} = \sup_{\pi \in I} \{\alpha'_\pi\}. \quad (45)$$

The value of the nuisance parameter to be used in establishing the critical value and test size is the value of π at which Equation 45 is maximized. In the next section we will demonstrate how this is done examining the same initial sample sizes as in Chapter II, $N_I(\cdot, \cdot) = 24, 48, 96$.

In summary, using the unconditional approach, we are not focusing on the values of $\pi_1 = \pi_2 = \pi$, but freeing the analysis of any specific value by reporting the results in which the test size is maximized. In the conditional approach, we are also trying to free the analysis of any specific value of $\pi_1 = \pi_2 = \pi$. Specifically, under the null, in the conditional approach, we make use of the hypergeometric distribution for the probability calculations. By doing this, we have freed the analysis of specific value of $\pi_1 = \pi_2 = \pi$. Unfortunately, under the alternative, we must specify the values of π_1 and π_2 . In the unconditional approach, we find the value of the nuisance parameter, $\pi_1 = \pi_2 = \pi$, that gives us a worst case scenario, i.e. it maximizes the overall test size across all possible final sample sizes. The appropriate nuisance parameter value is determined by the maximum test size value across the range of possible nuisance parameter values.

We realize that this approach is flawed. Since many different final sample sizes are likely even with a small initial sample size, e.g. for an initial sample size of 24, this approach will not produce the maximum test size value for any specific final sample

size. This approach does, however, produce the maximum expected test size across all final sample sizes.

3.2.2 Example of Finding Nuisance Parameter Values

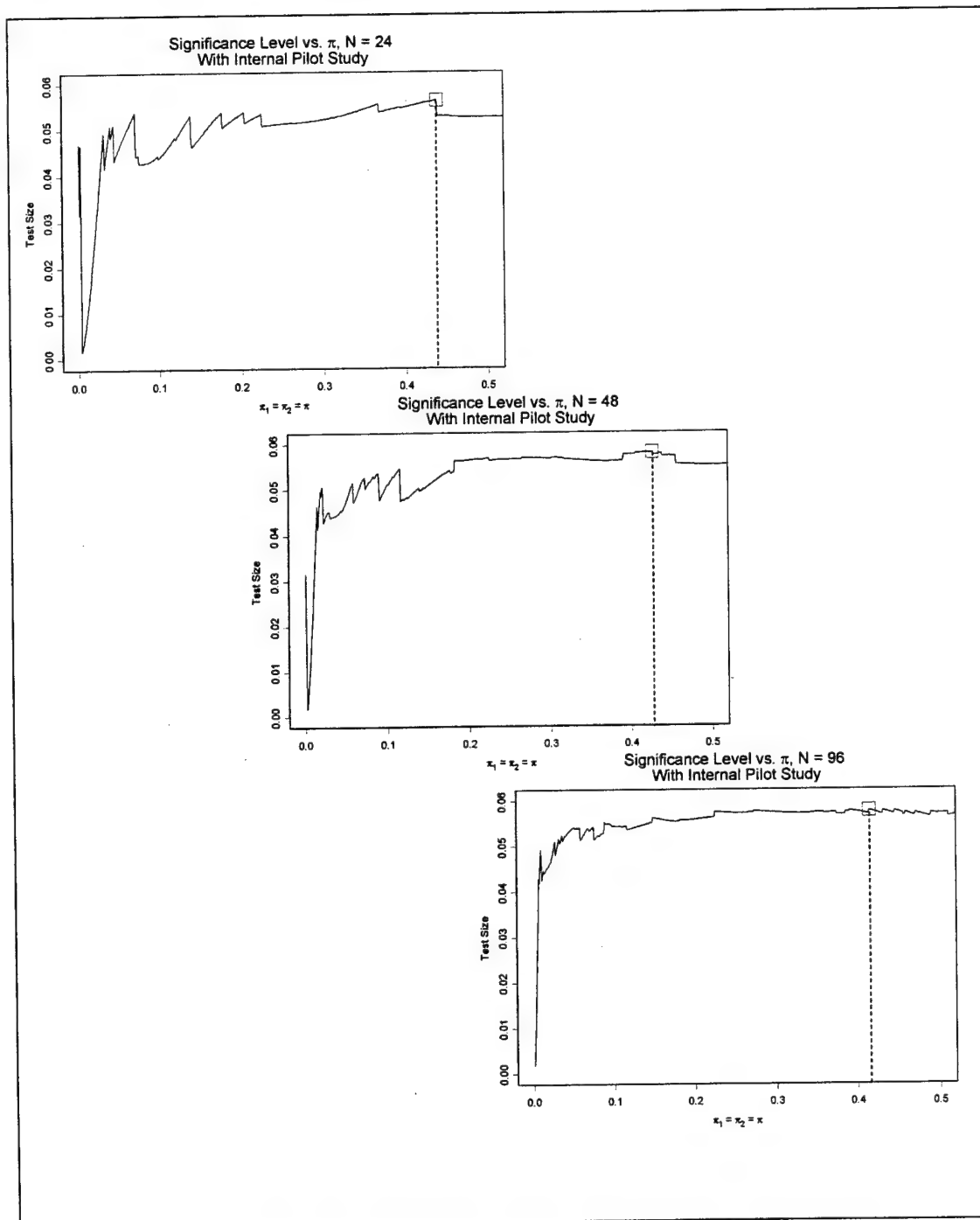
In this section, we develop a method for finding the appropriate nuisance parameter value. To do this, we follow the Suissa and Shuster method of using a grid search over the range 0.001 to 0.999 by 0.001 and examine the resulting probability of rejecting $H_0: \pi_1 = \pi_2 = \pi$, i.e. the test size. In reality, however, we need to only examine the values between 0.001 and 0.5 since the values above 0.5 are the mirror image of those below.

As stated above, the manner in which the test size is determined is analogous in either the conditional or unconditional approach. However, in the unconditional approach, we must maximize the test size over many possible values of the nuisance parameter, π . Figure 19 provides the results of examining test size for the initial sample sizes of 24, 48, and 96 by varying π .

We can determine the appropriate value of the nuisance parameter to use for each initial sample size, $N_I(\cdot, \cdot) = 24, 48, 96$. We find the value of $\pi_1 = \pi_2 = \pi$ such that the test size is the largest possible value, marked with a box in Figure 19. Notice that for different initial sample sizes, e.g. 48 and 96, the maximum value changes and occurs at a different value of the nuisance parameter. Table 7 provides the nuisance parameter values for initial sample sizes of 24, 48, and 96 that correspond to the appropriate box in Figure 19. We then use these values of the nuisance parameter to obtain critical values and create the power curves for each sample size.

Under the null, the value of the nuisance parameter corresponding to the largest test size value is used to calculate the internal pilot study, additional, and final 2×2 table probabilities. Under the alternative hypothesis, the value of the nuisance

parameter is equal to the proportion of controls with the genetic polymorphism of interest, $\pi_2 = \pi$, and π_1 is determined by the difference of interest, δ .



**Figure 19. Test Size Values for Initial Sample Sizes = 24, 48, 96.
Maximum value is highlighted with a box, "□".**

Initial Sample Size	$\pi = \pi_1 = \pi_2$	Test Size
24	0.438	0.05618
48	0.427	0.05798
96	0.415	0.05767

Table 7. Values of Nuisance Parameter, π , for Initial Sample Sizes of 24, 48, and 96.

3.2.3 Power and How to Use the Nuisance Parameter

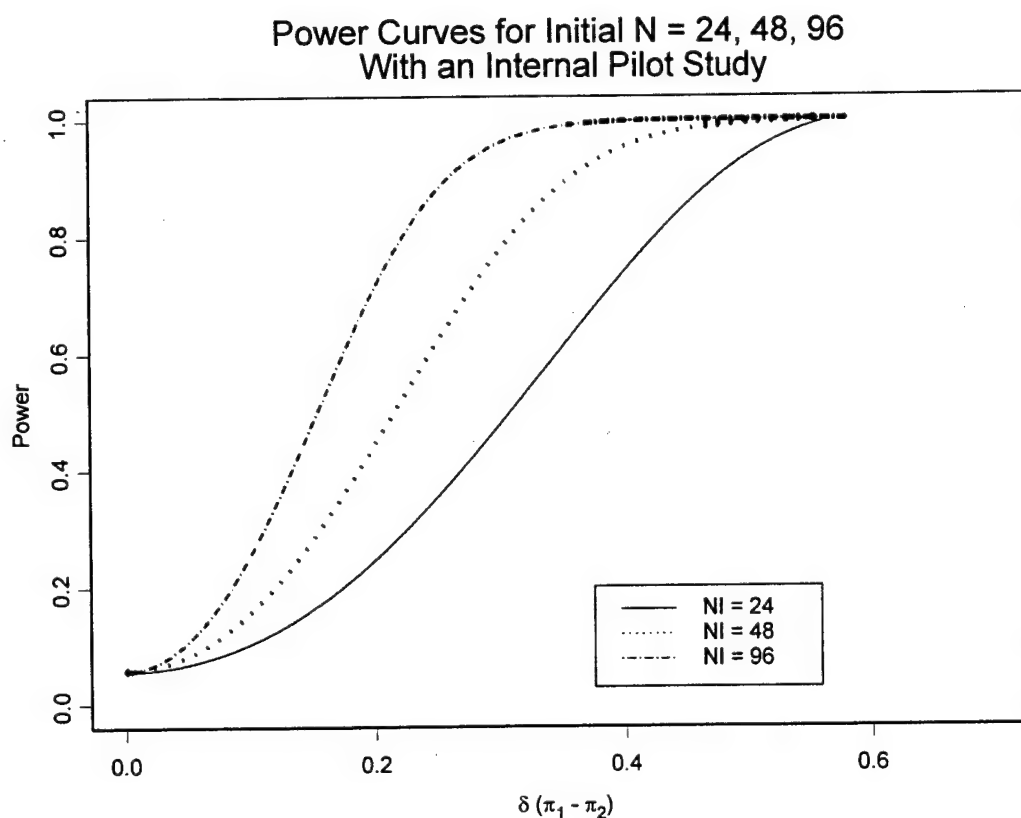
In this section, we use the nuisance parameter values in order to calculate the power and determine the appropriate critical values. The critical values are used, as in Chapter II, to determine the power. Making use of Theorem 2, § 2.2.3, we have a general form for calculating the power. As before, we first need to enumerate every possible table based on the internal pilot and additional sample sizes. Combining the internal pilot and additional sample sizes to get the final sample sizes, we can then decide whether to reject the null hypothesis or not based on the previously calculated critical values, the value of the nuisance parameter, and final sample size, $N_F(\cdot, \cdot)$. We can determine whether to reject each final table by comparing its test statistic value, T , against the associated critical value, T_{c_π} as in Equation 48. The power for the unconditional approach can then be calculated by summing the probabilities of those tables for which we reject the null hypothesis by varying the value of π_1 and fixing $\pi_2 = \pi$, Equation 46.

$$\text{Power} = \sum_{i=1}^B \sum_{j=1}^{M_i} \Pr\{T_{i,j} | \pi_1, \pi_2 = \pi\} \Pr\{T_{i,j} \in \text{Rejection Region} | \pi_1, \pi_2 = \pi\}. \quad (46)$$

Note that at the null, this equation is equivalent to Equation 44. As an example, we can create the power curves for the initial sample sizes of 24, 48, and 96 at the values of the

nuisance parameters given in Table 7. The power curves are provided below in Figure 20.

We have now developed the process to determine the appropriate nuisance parameter values, critical values, and power. However, we have not yet addressed the issue of test size inflation. Based on the results so far, we can show that the test size is inflated and to what degree when using an internal pilot study to re-estimate the final sample size using an unconditional approach.



Note: We are limiting the δ , $(\pi_1 - \pi_2)$, to positive values.

Figure 20. Power Curve for N = 24, 48, and 96.

Note that at a $\delta = 0$, the three values are not identical. Table 7 provides the value for each curve.

3.2.4 Test Size Inflation

As stated previously, the test size inflation can be determined by subtracting the discrete significance level, α_d , from the observed test size, α' . We have a slightly different notation to indicate that these values are actually maximized as in § 3.2.1. The notation used is $\alpha_{d_{max}}$ and α'_{max} . Where $\alpha_{d_{max}}$ is defined in Suissa and Shuster (1985) to be the maximum value attained without using an internal pilot study. The results are provided in Table 8.

Initial Sample Size	α'_{max}	$\alpha_{d_{max}}$	$\alpha'_{max} - \alpha_{d_{max}}$
24	0.05618	0.04997	0.00622
48	0.05798	0.04999	0.00799
96	0.05767	0.05000	0.00767

Table 8. Test Size Inflation with an IPS, Initial Sample Size = 24, 48, 96.
Test size inflation calculated by subtracting the significance level without an internal pilot study, α_d , and the significance level with an internal pilot study, α' .

In the unconditional approach, we not only have inflation, but inflation above the nominal significance level even with small sample sizes. Recall, the inflation is caused by the dependence of the final sample size on the internal pilot study data. What causes this inflation above the nominal significance level is the fact that in the unconditional approach the distribution of the final test statistic is less discrete. Since the test statistic is more discrete in the conditional case, the discrete significance level (α_d) is much smaller than the nominal significance level, α . There, since α_d is so much smaller than α (by as much as 0.02), even though there is inflation due to using an internal pilot study, the resulting test size, α' , is still below the nominal. By contrast, in the unconditional approach, the discrete significance level, $\alpha_{d_{max}}$, is very close to the nominal significance level; for example see Figure 8. Since $\alpha_{d_{max}}$ is so close to α in the unconditional approach, the resulting test size using an internal pilot study is above α . This result was not surprising and even expected.

3.2.5 Adjusting for Test Size Inflation

Adjusting for test size inflation is much more important in the unconditional test since inflation above the nominal significance level occurs at a much smaller initial sample size. In this section, we discuss what to do when the test size is inflated from using an internal pilot study while using an unconditional test. We use a binary type of search algorithm in order to adjust the maximum expected test size to be below the pre-specified nominal significance level. In this instance, since the test size inflation is much greater than in the conditional approach, it does not make sense to drop the nominal significance level by increments of 0.001 since it is more likely that the adjusted nominal significance level is much lower than the original level. The approach we use here is to drop the nominal significance level by 0.005 until the resulting test size inflation has dropped below the nominal level. Once this has occurred, we continue with the binary type search algorithm since we have bound the adjusted significance level. Table 9 provides the results of this procedure for an initial sample size of 24 in the order in which α was adjusted. We can see from Table 9 that to adjust for the test size inflation for an initial sample size of 24, we must use an adjusted nominal significance level of 0.042.

Nominal Significance Level, α	Discrete Significance Level, $\alpha_{d_{max}}$
0.05	0.0562
0.045	0.0528
0.04	0.0463
0.042	0.0495
0.043	0.0502

Table 9. Example of Adjustment Procedure at the Null, $\pi_1 = \pi_2 = \pi$, Initial Sample Size is 24. Where $\alpha_d = 0.04997$.

An additional complication for the unconditional test is that not only do we need to adjust the nominal significance level, we also need to adjust the value of the nuisance

parameter each time we adjust the value of the nominal significance level. For example, the appropriate value of the nuisance parameter at an adjusted nominal significance level of 0.042 is $\pi = 0.311$ versus $\pi = 0.438$ at a nominal significance level of 0.05. The reason this is mentioned is that if this is not done the adjustment process will be flawed. The point is that we are using the worse case scenario to free the unconditional analysis from any specific value of the nuisance parameter. To do this, we must follow the same process as we did under the null. That is, we must maximize the test size with relation to the value of the nuisance parameter.

3.3 Summary and Conclusions

If the analyst prefers an unconditional approach, we have now provided a means of analyzing a case-control study with an internal pilot study using an unconditional approach. This approach falls more in line with the type of data usually found in a case control study, that is, a fixed number of cases and controls with the number of cases and controls with or without the genetic polymorphism of interest allowed to vary.

Our approach is consistent with the unconditional approach of Suissa and Shuster (1985) in which they did not entertain the idea of an internal pilot. The technique used by Suissa and Shuster, since they did not use an internal pilot, did not address the concern of test size inflation. By using an internal pilot study, we have introduced the possibility of test size inflation. In the case of the unconditional approach, the test size inflation becomes a concern at a much smaller initial sample size than in the conditional approach. The reason for this is that the unconditional approach produces a final test statistic distribution which is less discrete than that for the conditional approach.

Now, this new approach to using an unconditional test also allows the analyst to adjust for the test size inflation caused by using an internal pilot study. This is an exact

method which also allows us to produce exact p -values since we have enumerated every possible 2×2 table and their probability of occurring. We can then determine the p -value based on summing the probabilities of the observed 2×2 table and those more extreme.

CHAPTER IV

MOTIVATING EXAMPLE

In this chapter, we provide an example based on the design and data from Marlar and Welsh (2001). We show how an applied scientist could use these methods in a grant application, during a study with re-estimation, and for data analysis. First, we compute the power, critical values, and required sample sizes using only the initial population parameter estimates of π_1 and π_2 . Next, we sample to create the internal pilot study data. Then, we calculate the required additional sample size based on an observed internal pilot study. Finally, we draw conclusions based on a set of observed final data. Since we do not have a complete experiment available from Marlar and Welsh and to simplify the example, we will step through a small fictitious example. However, we do make use of actual data from Marlar and Welsh from which we sample (with replacement) to come up with the observed data. Figure 21 provides a summary of the available data from Marlar and Welsh.

		Genetic Marker Factor V Leiden		
		+	-	
Cases	28	146		174
Controls	19	310		329
	47	456		503

Figure 21. 2×2 Contingency Table of Available Marlar and Welsh Data
This data is for the factor V Leiden genetic polymorphism only.

We will demonstrate both conditional and unconditional tests using the fictitious data sampled from the Marlar and Welsh data. We are assuming that half of the initial sample size estimate will be used for the internal pilot study, there are exactly

2 controls per case, the nominal significance level is 0.05, the planned power is 90%, the final sample size will never be smaller than initially estimated, and the final sample size will never be larger than twice that initially estimated.

For both the conditional and unconditional tests, we will first assume we have no data at all and only an estimate of the population parameters π_1 and π_2 . Next, we assume that we only have data from the internal pilot. Finally, we will assume that we have the final/additional data and actually make a decision regarding the association of Factor V Leiden with VTE based on these data.

To begin a study, we must first have some kind of hypothesis we would like to test. In this case, we are hypothesizing that there is a difference in the proportion of cases and controls with the genetic polymorphism of interest; $H_0: \pi_1 = \pi_2$ and $H_1: \pi_1 \neq \pi_2$. For this fictitious example, we are assuming that the estimates of the population parameters are 0.4 and 0.1 for the cases and controls, respectively. Specifically, we are interested in whether the cases have a different proportion with the genetic polymorphism than the controls. Then, we must obtain estimates of the population parameters. In this instance, the parameters of interest are the proportion of cases and controls with the genetic polymorphism, π_1 and π_2 . The initial estimates are assumed to have come from the current literature. Note, we will use the Rosner (1995) sample size equation for a difference in two proportions for all of the sample size calculations in this chapter.

With the population parameter estimates in hand, we are now ready to obtain an initial sample size estimate, $N_I(\cdot, \cdot)$. Using the proportions from above, the required initial sample size, $N_I(\cdot, \cdot)$, to maintain a power of 90% is 93 subjects. Recall, the internal pilot study sample size, $N_P(\cdot, \cdot)$, is half the initial sample size estimate. Therefore, the internal pilot study sample size is 93/2 or 46.5 subjects. However, we cannot have half a subject and must round up to the nearest whole subject that is

divisible by 3. The reason the internal pilot study sample size has to be divisible by three is because of the ratio of cases to controls - the number of subjects has to be divisible by 3 in order to maintain the 1 : 2 ratio. Given this, the internal pilot study sample size, $N_P(\cdot, \cdot)$, is 48. Based on the 1 : 2 ratio, we have $N_P(1, \cdot) = 16$ cases and $N_P(2, \cdot) = 32$ controls which will be used in the examination of both the conditional and unconditional tests. In § 4.1, we will discuss the conditional test and in § 4.2, we will discuss the unconditional test.

4.1 Exact Conditional Test

In this section, we develop the power and critical values necessary for the initial sample size calculated above for the conditional test. Then, we randomly sample (with replacement) from the Marlar and Welsh data to develop the fictitious example dataset. Using the power and critical values developed in this section, we then show the practitioner how to make the appropriate conclusion using the randomly generated data.

4.1.1 *What We Would do for a Grant Proposal*

Now that we know the internal pilot study sample size, $N_P(\cdot, \cdot) = 48$, we can enumerate all possible internal pilot study 2×2 tables. From these internal pilot study tables, we can estimate the population parameters and recalculate the required final sample size, $N_F(\cdot, \cdot)$. The difference between the final sample size and the internal pilot study sample size is the additional sample size, $N_A(\cdot, \cdot)$. Now, from each of the enumerated internal pilot study tables, we can enumerate all possible additional tables. By combining the internal pilot study and additional data, we have all possible final tables. By calculating the probabilities of all the final tables and their associated test statistic values, we can determine the distribution of the final test statistic as in § 2.2.1 and § 2.2.2. Using the distribution of the final test statistic along with the critical values (§ 2.2.3), we can determine the expected power under the null and alternative

hypotheses. We then have the following power curve derived from §2.2.1 - §2.2.3, Figure 22.

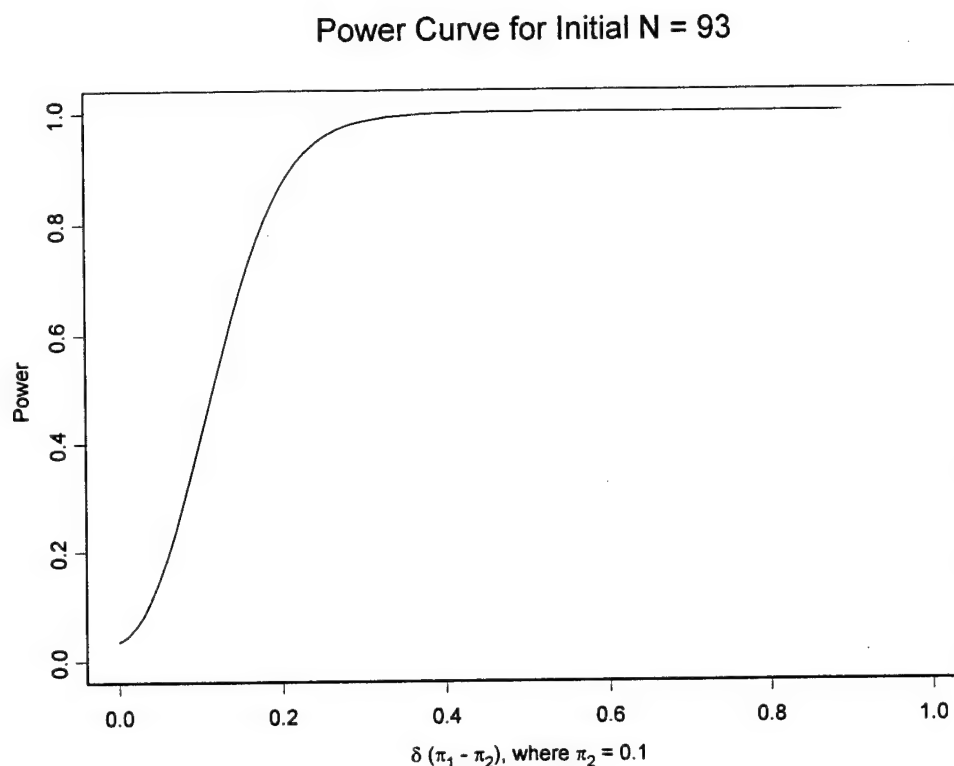


Figure 22. Power Curve for $N_I(\cdot, \cdot) = 93$, Conditional Approach.

As you can see from Figure 22, the expected power is somewhat higher than anticipated. This is an artifact of the settings used. Recall that the internal pilot study is only used to re-estimate the final sample size. The rules placed on the sample size re-estimation are that the minimum final sample size will never be below that initially estimated and the maximum final sample size will never be more than twice that initially estimated. Therefore, when we observe an internal pilot study in which the proportions are approximately equal, we have a reestimated final sample size at twice the initial (the maximum allowed). This in turn, through the enumeration process, leads

to a number of final tables that we can in fact reject although they were spawned from an internal pilot study table in which rejection looked implausible.

4.1.2 Midterm Analysis - Sample Size Re-estimation

Now, in order to actually create the internal pilot study data, we will need to sample, with replacement, from the Marlar and Welsh data, Figure 21. Figure 23 provides the results of the random sampling from the Marlar and Welsh data to simulate observing an internal pilot study.

Internal Pilot Study Data:

	Genetic Marker		
	+	-	
Cases	5	11	16
Controls	2	30	32
	7	41	48

Figure 23. Internal Pilot Study 2×2 Table from Random Experiment

Now that we have observed internal pilot study data, using the Rosner sample size equation once again, we can recalculate the required final sample size. Our new population parameter estimates are $\pi_1 = 5/16$ and $\pi_2 = 2/32$ for cases and controls respectively. We now have a recalculated final sample size of $N_F(\cdot, \cdot) = 108$. Although the parameter estimates for π_1 and π_2 have decreased, which intuitively decreases the variance, the sample size increased. The reason for this is that the sample size equation also contains the observed difference which increases the required sample size by more than any savings from the decreased variance. In this case, the final sample size has not been affected by the assumptions: 1) minimum sample size of $N_I(\cdot, \cdot)$, and 2) maximum sample size of $2N_I(\cdot, \cdot)$. Since we have already collected data on 48 subjects, we need only collect data on an additional 60 subjects, $N_A(\cdot, \cdot)$.

4.1.3 Final Data Analysis and Conclusions

In order to finish with the data collection, we need to randomly sample once again from the Marlar and Welsh data in order to simulate the additional data needed. Finally, let us pretend that we have observed the additional data and hence the final 2×2 table, Figure 24, and let us make some inference.

<u>Additional Data:</u>				<u>Final Data:</u>					
		Genetic Marker				Genetic Marker			
		+	-			+	-		
Cases		6	14	20	Cases	11	25	36	
Controls		2	38	40	Controls	4	68	72	
		8	52	60			15	93	108

Figure 24. Additional and Final 2×2 Tables from Random Experiment

From the final 2×2 table in Figure 24, we have a calculated final test statistic value, T , of 12.5419 which comes from the test statistic equation in § 1.3.6. Now that we have an observed test statistic value, we can compare it to the critical value, T_c , under the same conditions; specifically, $N_F(\cdot, \cdot) = 108$ and $N_F(\cdot, 1) = 15$. Recall from § 1.3.7 and § 2.2.3 that the critical value is defined as the largest test statistic value such that the probability that we reject the null hypothesis is at most α when the null hypothesis is in fact true. If the test statistic value, T , is greater than or equal to the critical value, there is evidence to reject the null hypothesis, $H_0: \pi_1 = \pi_2$. For these particular conditions, the critical value, T_c , is equal to 5.5742 with $\alpha_d = 0.0353$. Since $T \geq T_c$, the conclusion is that there is sufficient evidence to reject the null hypothesis or that there is sufficient evidence against $\pi_1 = \pi_2$. What this means to the practitioner is that we have evidence that the proportion of cases with the genetic polymorphism of interest is different than the proportion of controls with the genetic polymorphism of

interest. This is evidence that the genetic polymorphism of interest is related to venous thrombosis.

Another approach that could have been used to determine the significance of the observed final table is through using the p -value. In this instance, the p -value is the probability of observing a table this or more extreme assuming that the null hypothesis is true, where $H_0 : \pi_1 = \pi_2 = 0.1$. To calculate the p -value, we need to sum the probabilities of the observed 2×2 table and of those more extreme tables. For the observed final table in Figure 21, the p -value is 6.115×10^{-6} . Comparing this value against either the nominal significance level of 0.05 or the discrete significance level of 0.0353, it is clearly significant.

4.1.4 New Sample Size Methodology

As part of the grant writing process, we should examine the power at the difference of interest hypothesized by the investigators. The power at $\pi_1 = 0.4$ and $\pi_2 = 0.1$, or $\delta = 0.3$, is equal to 0.9846. Recall, the over-powered situation is caused by the limiting conditions as discussed in § 4.1.1. Since the power is so high at the hypothesized difference using the Rosner (1995) sample size equation, we thought that this would be a good opportunity to demonstrate the proposed sample size approach from § 2.2.6. It is presented here as an example only and not used in the remaining sections of this chapter.

Since an initial sample size of 93 produced a power greater than planned, this translates into a sample size which is too large. Using the methods introduced in § 2.2.6, we can converge on a sample size which produces a power within a pre-specified tolerance of that planned. Table 10 provides the results of this process.

Initial Sample Size	Power @ $\pi_1 = 0.1$ and $\pi_2 = 0.4$
93	0.9846
48	0.8684
72	0.957
60	0.926
54	0.9003

Table 10. Results from Improved Sample Size Methodology.

4.2 Exact Unconditional Test

In this section, we obtain the power and critical values necessary for the initial sample size calculated above for the unconditional test. We make use of the same randomly sampled fictitious example data from § 4.1.2, Figures 20 and 21. Using the power and critical values developed in this section, we then show the practitioner how to make the appropriate conclusion using the randomly generated data.

4.2.1 What We Would do for a Grant Proposal

For the unconditional approach, we will use the same internal pilot study sample size, $N_P(\cdot, \cdot) = 48$, the same additional sample size, $N_A(\cdot, \cdot) = 60$, and hence the same recalculated final sample size, $N_F(\cdot, \cdot) = 108$. Now, we have the same enumerated internal pilot study and additional tables as with the conditional approach and therefore the same test statistic values. By calculating the probabilities of all the final tables and their associated test statistic values across the range of nuisance parameters, we can determine the appropriate value of the nuisance parameter and then the distribution of the final test statistic as in § 3.2.1 - § 3.2.3. For an initial sample size of 93, the value of the nuisance parameter, π , that maximizes the test size is 0.397. Note that $\alpha'_{max} = 0.0576$ is inflated above the nominal level, $\alpha = 0.05$. We can similarly determine the power under any alternative hypothesis by maximizing the test size as done under the null. Note that the power at $\delta = 0.3$, a meaningful difference for the conditional test, is equal to 0.9999.

As with the conditional approach, the unconditional approach has an expected power somewhat higher than anticipated. Again, this is an artifact of the settings used which increases the power dramatically.

Before we proceed, we must examine the test size inflation. Recall the value of α'_{max} , 0.0576, which is not only inflated, but inflated above the nominal significance level of 0.05. The maximum discrete significance level, $\alpha_{d_{max}}$, was found to be 0.05 to 6 decimal places. In order to adjust for the test size inflation, we must use the procedure set forth in § 3.2.5. In using this procedure, we found that adjusting the critical value calculations using a nominal significance level of 0.042, adjusts the test size to be less than or equal to the nominal significance level of 0.05. Therefore, these critical values and $\alpha_{d_{max}}$ values should be used in the final analysis to determine whether a table is significant or not.

Now, since we know we have to adjust for test size inflation above the nominal significance level, we will also need to use these values to determine the power under any alternative. Using the same procedure as above, we can determine the expected power under the null and alternative hypotheses for the adjusted nominal significance level.

4.2.2 Midterm Analysis - Sample Size Re-estimation

Now, in order to actually create the internal pilot study data, we will need to sample, with replacement, from the Marlar and Welsh data, Figure 18. Figure 25 provides the results of the random sampling from the Marlar and Welsh data to simulate observing an internal pilot study. This is the same data as used for the conditional approach.

	Genetic Marker		
	+	-	
Cases	5	11	16
Controls	2	30	32
	7	41	48

Figure 25. Internal Pilot Study 2×2 Table from Random Experiment

Now that we have observed internal pilot study data, using the Rosner (1995) sample size equation once again, we can recalculate the required final sample size. Our new population parameter estimates are $\pi_1 = 5/16$ and $\pi_2 = 2/32$ for cases and controls respectively. Since we have not changed the sample size re-estimation procedure from the conditional approach, the final sample size of $N_F(\cdot, \cdot) = 108$ remains the same as does the additional subjects, $N_A(\cdot, \cdot) = 60$.

4.2.3 Final Data Analysis and Conclusions

Therefore, we need to randomly sample once again from the Marlar and Welsh data in order to simulate the additional data needed. Finally, let us pretend that we have observed the same additional data and hence the same final 2×2 table as in § 4.1.3, provided again in Figure 26 for convenience. What do we infer from these data?

<u>Additional Data:</u>				<u>Final Data:</u>			
	Genetic Marker				Genetic Marker		
	+	-			+	-	
Cases	6	14	20	Cases	11	25	36
Controls	2	38	40	Controls	4	68	72
	8	52	60		15	93	108

Figure 26. Additional and Final 2×2 Tables from Random Experiment

From the final 2×2 table in Figure 26, we have a calculated final test statistic value, T , of 12.5419 which comes from the test statistic equation in § 1.3.6. Now that we have an observed test statistic value, we can compare it to the critical value, $T_{c\pi}$,

under the same conditions; specifically, $N_F(\cdot, \cdot) = 108$ and the appropriate value of π . If the test statistic value, T , is greater than or equal to the critical value, there is evidence to reject the null hypothesis, $H_0: \pi_1 = \pi_2$. For these particular conditions, the critical value, T_{c_π} , is equal to 3.8571 with $\alpha_{d_{max}} = 0.04845$ assuming the null is true. Since $T \geq T_{c_\pi}$, the conclusion is that there is sufficient evidence to reject the null hypothesis or that there is sufficient evidence against $\pi_1 = \pi_2$. Therefore, we make the same conclusion as in the conditional case. There is evidence that the genetic polymorphism of interest is related to venous thrombosis.

As in the conditional approach, we can also assess the significance by means of the p -value. For the observed final table in Figure 28, the p -value is 1.266×10^{-15} . Comparing this value against either the nominal significance level of 0.05 or the discrete significance level of 0.04845, it is clearly significant.

4.3 Summary and Conclusions

For this particular example, the same conclusion was made regardless of the approach used, either conditional or unconditional. However, given the somewhat large difference in both the critical values and p -values, we could have a situation in which the conclusion could differ based on which approach was used, especially if the observed 2×2 table were not quite as extreme as the one we observed here.

Now, we have given the practitioner an example with which s/he can follow how the process should work for a particular problem. Depending on which approach is preferred, either the conditional or unconditional, both approaches are provided.

CHAPTER V

DISCUSSION

5.1 Summary and Conclusions

The idea of using an internal pilot study in sample size recalculation is very practical and beneficial. First, we get to re-estimate the parameters of interest without throwing out any data. Next, we are dramatically increasing the chances of a successful study by using parameter estimates from the correct population of interest. Finally, the use of an internal pilot can be applied to almost any type of study. Our focus however is on binary data from an observational study.

Binary data occur quite often in the area of biometrics and epidemiology. There is a large amount of literature addressing the issues concerning binary data. However, this is not the case with regard to binary data and sample size recalculation in the form of an internal pilot study. Several authors address the issue of sample size recalculation using an internal pilot study while using continuous outcome data, specifically, normally distributed data. In this case, the internal pilot study is used to re-estimate the variance. The final sample size is then recalculated using the new variance estimate and the same minimally detectable difference of interest. A side-effect of this process is an inflation in test size due to the dependence of the final sample size on the internal pilot study. The final sample size is now a random variable that is dependent on the internal pilot study data through the re-estimation of the variance. The good news with normally distributed data is that the mean and variance are independent. Therefore, we can find a form of the test statistic that is independent of the re-estimated variance and consequently account for the inflation in test size. This being said, the literature is comprehensive in the area of internal pilot study use with a normally distributed

outcome variable. This is not the case, however, with the use of a binary outcome variable. The literature is much less developed in this area.

What we have done is provide the analyst and practitioner with a means of using an internal pilot study with a binary outcome variable. Theorems 1 and 2 provide generic results to determine table probabilities, power, and significance levels for any categorical test. This approach provides a general means of determining the distribution of the final test statistic and hence making exact inference for a wide range of categorical data problems. We have provided both the conditional and unconditional approaches which cover the gamut of possible problems. Therefore, the analyst or practitioner can pick the approach that matches the assumptions made. In both the conditional and unconditional approaches, we can now calculate the appropriate critical values and p -values to perform the proper inference and adjust for the test size inflation when necessary. For the conditional approach with small samples, we get the benefit of reestimating the sample size without paying a penalty in terms of test size inflation above the nominal significance level. This is not true, however, for the unconditional approach. In the unconditional approach, even in small samples, we must adjust for test size inflation above the nominal. We realize that there has been and still is some controversy over the appropriateness of the unconditional approach. However, we believe that in some instances it is appropriate to use an unconditional approach based on the type of data and the assumptions made.

The primary limitation of this thesis is that we have examined the simplest form of the problem, i.e., data in the form of a 2×2 table assuming no interactions or confounders. We believe, however, that this is a reasonable first approach to this problem since it has not yet been addressed in the literature. We must get a good understanding of the simple problem before we can solve the larger more complicated one. Although our testing procedure is not uniformly most powerful nor based on

maximum likelihood for either the conditional or unconditional approaches, it is implementable and one of the first procedure developed for use with categorical data in observational studies.

This thesis provides one of the first pieces of literature addressing the major gap in this area of research, using an internal pilot study for sample size recalculation. Although not perfect, we have provided a means to account for the test size inflation when using an internal pilot study with binary data with either a conditional or unconditional test. The approaches within this thesis are a good initial cut at the problem and provide an excellent opportunity for future research.

5.2 Directions for Future Research

There are numerous directions in which this research could extend and a variety of ways of analyzing data with a binary outcome and exposure. We could examine different modeling approaches, different forms of the data, or change the model assumptions. We could also just make small extensions of the existing analysis or small changes in the data assumptions. Finally, we could examine how the results from this analysis could be applied to confidence intervals. For instance, should we change the way in which we calculate a confidence interval based on these results?

First, let us examine a possibly different modeling approach. Instead of using data in the form of a 2×2 table, we could use a logistic regression or an additive model approach. This would allow us to account for interactions and confounders. We could estimate the relative risk of exposure for important confounders, including plausible interactions between exposures.

Now, what about different forms of binary data? With binary data, we can examine the data in one of three ways: 1) a difference, such as $p_1 - p_2$, which would be of primary interest in a clinical trial, 2) a ratio, such as $\frac{p_1}{p_2}$, which would be of interest in

a cohort study, and finally 3) the odds ratio which would be of primary interest in a case-control study. We have only examined the data in terms of a difference.

Therefore, we can examine a ratio or odds ratio and adjust our approach to handle either of these two approaches.

Next, we can examine changes in the model assumptions. By changing the assumptions, we would change the problem dramatically. For example, if we assumed that the initial estimates are made with error, this error would be passed along to any analysis performed and add an additional degree of complexity to the problem. Also, we could assume that we will sample until we have an appropriate number of subjects with exposure. That is, we could use a negative binomial for the underlying distribution and probability calculations.

We could also examine variations in the secondary parameters. For instance, we could vary the percentage of the initial sample size to use as the internal pilot study. We could also change the minimum and maximum sample size requirements, i.e., a maximum not limited to twice the initial sample size estimate.

Finally, we could look into some issues that came from the original analysis. For example, in Chapter IV, we noticed that the conditional test was more powerful than the unconditional test at a δ of 0.3, i.e. 0.9846 versus 0.9569. We could examine the difference between the conditional and unconditional approaches and actually calculate how often we make a mistake in each case. For example, with an enumeration approach, we are examining every 2×2 table, which are identical for both approaches. Therefore, we can examine each individual 2×2 table under both the conditional and unconditional approach and determine whether we reject or accept each individual 2×2 table. Since we actually know what the truth is, we can determine how each performs.

BIBLIOGRAPHY

- Andersen, Per Kragh (1987). Conditional Power Calculations as an Aid in the Decision Whether to Continue a Clinical Trial. *Controlled Clinical Trials* **8**, 67-74
- Barnard, G. A. (1947). Significance test for 2×2 tables. *Biometrika* **34**, 123-138.
- Barnard, G. A. (1949). Statistical inference. *Journal of the Royal Statistical Society, Ser. B*, **11**, 115-139.
- Bartlett, M. S. (1953). Approximate confidence intervals. *Biometrika* **40**, 306-317.
- Betensky, R. A. and C. Tierney (1997). An examination of methods for sample size recalculation during an experiment. *Statistics in Medicine* **16**, 2587-2598.
- Birkett, M. A. and S. J. Day (1994). Internal pilot studies for estimating sample size. *Statistics in Medicine* **13**, 2455-2463.
- Birnbaum, A. (1962). On the foundations of statistical inference (with discussion). *Journal of Statistical Planning and Inference* **57**, 269-326.
- Browne, R. H. (1995). On the use of a pilot sample for sample size determination. *Statistics in Medicine* **14**, 1933-1940.
- Chen, V. C. P. and A. J. Hayter (1996). Sensitivity analysis of upper confidence bounds on the range of treatment effects. *Computational Statistics & Data Analysis* **23**, 257-262.
- Coffey, C. S. and K. E. Muller (1999). Exact test size and power of a Gaussian error linear model for an internal pilot study. *Statistics in Medicine* **18**, 1199-1214.
- Coffey, C. S. and K. E. Muller (2001). Controlling test size while gaining the benefits of an internal pilot design. *Biometrics* **57**, 625-631.
- Cordeiro, G. M. (1983). Improved likelihood ratio statistics for generalized linear models. *Journal of the Royal Statistical Society, Series B* **45**, 404-413.
- Cytel Software Corporation (1999). Proc-StatXact 4 For SAS Users, Users Manual.
- Davies, R. B. (1980). The distribution of a linear combination of χ^2 random variables. *Applied Statistics* 323-333.
- Day, S. (2000). Operational difficulties with internal pilot studies to update sample size. *Drug Information Journal* **34**, 461-468.

- Denne, J. S. (2001). Sample size recalculation using conditional power. *Statistics in Medicine* **20**, 2645-2660.
- DiSantostefano, R. L. and K. E. Muller (1995). A comparison of power and approximations for Satterthwaite's test. *Communications in Statistics - Simulation and Computation* **24**(3), 583-593.
- Dozier, W. G. and K. E. Muller (1993). Small-sample power of uncorrected and Satterthwaite corrected *t*-tests for comparing binomial proportions. *Communications in Statistics - Simulation and Computation* **22**(1), 245-264.
- Everitt, B. S. The Analysis of Contingency Tables. Chapman and Hall, 1977.
- Foppa, I. and D. Spiegelman (1997). Power and sample size calculations for case-control studies of gene-environment interactions with a polytomous exposure variable. *American Journal of Epidemiology* **146**, 596-604.
- Gordis, Leon. *Epidemiology*. W. B. Saunders Company, 1996; 124-140.
- Gould, A. L. (1983). Sample sizes required for binomial trials when the true response rates are estimated. *Journal of Statistical Planning and Inference* **8**, 51-58.
- Gould, A. L. (1992). Interim analyses for monitoring clinical trials that do not materially affect the type I error rate. *Statistics in Medicine* **11**, 55-66.
- Gould, A. L. (1995). Planning and revising the sample size for a trial. *Statistics in Medicine* **14**, 1039-1051.
- Gould, A. L. (2001). Sample size re-estimation: Recent developments and practical considerations. *Statistics in Medicine* **20**, 2625-2643.
- Greenland, S. (1988). On sample-size and power calculations for studies using confidence intervals. *American Journal of Epidemiology* **128**, 231-237.
- Herson, J. and J. Wittes (1993). The use of interim analysis for sample size adjustment. *Drug Information Journal* **27**, 753-760.
- Jennison, C. and B. W. Turnbull (2000). *Group Sequential Methods with Applications to Clinical Trials*. Boca Raton, Florida; Chapman and Hall/CRC.
- Keiser, M. and T. Friede (2000a). Re-calculating the sample in internal pilot study designs with control of the type I error rate. *Statistics in Medicine* **19**, 901-911.
- Keiser, M. and T. Friede (2000b). Blinded sample size reestimation in multiarmed clinical trials. *Drug Information Journal* **34**, 455-460.

- Lachin, J. M. (1977). Sample size determination for $r \times c$ comparative trials. *Biometrics* **33**, 315-324.
- Lachin, J. M. (1981). Introduction to sample size determination and power analysis for clinical trials. *Controlled Clinical Trials* **2**, 93-113.
- Laird, N. M., M. C. Weistein and W. B. Stason (1979). Sample-size estimation: A sensitivity analysis in the context of a clinical trial for treatment of mild hypertension. *American Journal of Epidemiology* **109**, 408-419.
- Law, M. G. (1996). Sample size calculations for within-patient comparisons with a binary or survival endpoint. *Controlled Clinical Trials* **17**, 221-225.
- Lawley, D. N. (1956). A general method for approximating to the distribution of likelihood ratio criteria. *Biometrika* **43**, 295-303.
- Little, Roderick J. A. (1989). Testing the equality of two independent binomial proportions. *American Statistical Association* **43**, 283-288.
- Marlar, R. A., and C. H. Welsh (1998). Epidemiology of genetic and acquired risk factors for venous thrombosis.
- Mood, A. M., F. A. Graybill, and D. C. Boes. *Introduction to the Theory of Statistics*, 3rd Edition. McGraw-Hill, Inc., 1974.
- Muller, K. E. and V. B. Pasour (1997). Bias in linear model power and sample size due to estimating variance. *Communications in Statistics - Theory and Methods* **26**(4), 839-851.
- Peace, K. E. (1993). Discussion for interim analysis and sample size reestimation. *Drug Information Journal* **27**, 765-769.
- Posh, M. and P. Bauer (2000). Interim analysis and sample size reassessment. *Biometrics* **56**, 1170-1176.
- Proschan, M. A. (2000). An improved double sampling procedure based on the variance. *Biometrics* **56**, 1183-1187.
- Qiu, P. *et al.* (2000). Sample size to test for interaction between a specific exposure and a second risk factor in a pair-matched case-control study. *Statistics in Medicine* **19**, 923-935.
- Rosner, B. Fundamentals of Biostatistics, 4th Edition. Brooks/Cole Publishing Company, 1995; 383-386.

- Sandvik, L. et al. (1996). A method for determining the size of internal pilot studies. *Statistics in Medicine* **15**, 1587-1590.
- Satten, G. A. and L. L. Kupper (1990). Sample size requirements for interval estimation of the odds ratio. *American Journal of Epidemiology* **131**, 177-184.
- Schlesselman, J. J. (1974). Sample size requirements in cohort and case-control studies of disease. *American Journal of Epidemiology* **99**, 381-384.
- Self, S. G. and R. Mauritsen (1988). Power/sample size calculations for generalized linear models. *Biometrics* **44**, 79-86.
- Self, S. G., R. Mauritsen, and J. Ohara (1992). Power calculations for likelihood ratio tests in generalized linear models. *Biometrics* **48**, 31-39.
- Selicato, G. R. and K. E. Muller (1998). Approximating power of the unconditional test for correlated binary pairs. *Communications in Statistics -- Simulation and Computation* **27**(2), 553-564.
- Schlesselman, J. J. (1982). *Case-Control Studies: Design, Conduct, Analysis*. Oxford University Press, New York.
- Shieh, G. (2000). On power and sample size calculations for likelihood ratio tests in generalized linear models. *Biometrics* **56**, 1192-1196.
- Shih, W. J. (1993). Sample size reestimation for triple blind clinical trials. *Drug Information Journal* **27**, 761-764.
- Shih, W. J. and P. Zhao (1997). Design for sample size re-estimation with interim data for double-blind clinical trials with binary outcomes. *Statistics in Medicine* **16**, 1913-1923.
- Spiegelhalter, D. J. et al. (1986). Monitoring clinical trials: conditional or predictive power? *Controlled Clinical Trials* **7**, 8-17
- Stein, C. (1945). A two-sample test for a linear hypothesis whose power is independent of the variance. *Annals of Mathematical Statistics* **16**, 243-258.
- Stromberg, U. (1997). A method for deciding early stopping of inconclusive case-control studies in settings where data are stratified. *Statistics in Medicine* **16**, 2327-2337.
- Suissa, S. and J. J. Shuster (1985). Exact Unconditional Samples Sizes for the 2×2 Binomial Trial. *Journal of the Royal Statistical Society, Ser. A*, **148**, 317-327.

- Suissa, S. and J. J. Shuster (1991). The 2×2 Matched-Pairs Trial: Exact Unconditional Design and Analysis. *Biometrics* **47**, 361-372.
- Taylor, D. J. and K. E. Muller (1996). Bias in linear model power and sample size calculation due to estimating noncentrality. *Communications in Statistics - Theory and Methods* **25**(7), 1595-1610.
- Yates, F. (1984). Tests of Significance for 2×2 Contingency Tables (with discussion). *Journal of the Royal Statistical Society, Ser. A*, **147**, 426-463.
- Wittes, J. and E. Brittain (1990). The role of internal pilot studies in increasing the efficiency of clinical trials. *Statistics in Medicine* **9**, 65-72.
- Wittes, J. *et al.* (1999). Internal pilot studies I: Type I error rate of the naive *t*-test. *Statistics in Medicine* **18**, 3481-3491.
- Zucker, D. M. *et al.* (1999). Internal pilot studies II: Comparison of various procedures. *Statistics in Medicine* **18**, 3493-3509.

Appendix A

Notation and Definitions

Notation	Definition
π_1	Population Proportion of Cases with Genetic Marker
π_2	Population Proportion of Controls with Genetic Marker
a	Percentage of Initial Sample to Use for Internal Pilot Study
I	Initial
P	Internal Pilot Study
F	Final
k	Ratio of Controls to Cases
$N_I(1, \cdot)$	Calculated Initial Sample Size for Cases
$N_I(2, \cdot)$	Calculated Initial Sample Size for Controls
$N_I(\cdot, \cdot)$	Calculated Initial Total Sample Size
$n_P(1, 1)$	Observed Number of Cases with Genetic Marker in Internal Pilot Study
$n_P(1, 2)$	Observed Number of Cases without Genetic Marker in Internal Pilot Study
$N_P(1, \cdot)$	Calculated Internal Pilot Study Sample Size for Cases
$n_P(2, 1)$	Observed Number of Controls with Genetic Marker in Internal Pilot Study
$n_P(2, 2)$	Observed Number of Controls without Genetic Marker in Internal Pilot Study
$N_P(2, \cdot)$	Calculated Internal Pilot Study Sample Size for Controls
$N_P(\cdot, \cdot)$	Calculated Internal Pilot Study Total Sample Size
$n_A(1, 1)$	Observed Number of Cases with Genetic Marker in Additional Sample Size
$n_A(1, 2)$	Observed Number of Cases without Genetic Marker in Additional Sample Size
$N_A(1, \cdot)$	Calculated Additional Sample Size for Cases
$n_A(2, 1)$	Observed Number of Controls with Genetic Marker in Additional Sample Size
$n_A(2, 2)$	Observed Number of Controls without Genetic Marker in Additional Sample Size
$N_A(2, \cdot)$	Calculated Additional Sample Size for Controls
$N_A(\cdot, \cdot)$	Calculated Additional Total Sample Size
$n_F(1, 1)$	Observed Number of Cases with Genetic Marker in Final Sample Size
$n_F(1, 2)$	Observed Number of Cases without Genetic Marker in Final Sample Size
$N_F(1, \cdot)$	Calculated Final Sample Size for Cases
$n_F(2, 1)$	Observed Number of Controls with Genetic Marker in Final Sample Size
$n_F(2, 2)$	Observed Number of Controls without Genetic Marker in Final Sample Size
$N_F(2, \cdot)$	Calculated Final Sample Size for Controls
$N_F(\cdot, \cdot)$	Calculated Final Total Sample Size
$p_I(1)$	Proportion of Cases with Genetic Marker (Initial Estimate)
$p_I(2)$	Proportion of Controls with Genetic Marker (Initial Estimate)
$p_P(1)$	Proportion of Cases with Genetic Marker from Internal Pilot Study
$p_P(2)$	Proportion of Controls with Genetic Marker from Internal Pilot Study
$p_F(1)$	Proportion of Cases with Genetic Marker from Final Sample
$p_F(2)$	Proportion of Controls with Genetic Marker from Final Sample
α	Nominal Significance Level
α_d	Discrete Significance Level
$\alpha_{d_{max}}$	Maximum Discrete Significance Level from Unconditional Approach
α'	Test Size
α'_{max}	Maximum Test Size from Unconditional Approach
$1 - \beta$	Power

Appendix B

More Complete Table of Significance Levels

		0.05	0.1	0.15	0.2	0.25	π	0.3	0.35	0.4	0.45	0.5
	24	8.03×10^{-3}	1.38×10^{-2}	1.61×10^{-2}	1.55×10^{-2}	1.33×10^{-2}	1.12×10^{-2}	9.57×10^{-3}	8.77×10^{-3}	8.66×10^{-3}	8.72×10^{-3}	
N	48	1.42×10^{-2}	1.04×10^{-2}	8.39×10^{-3}	9.72×10^{-3}	1.05×10^{-2}	1.28×10^{-2}	1.56×10^{-2}	1.48×10^{-2}	9.53×10^{-3}	6.30×10^{-3}	
	96	6.18×10^{-3}	6.81×10^{-3}	6.98×10^{-3}	8.10×10^{-3}	9.24×10^{-3}	7.96×10^{-3}	5.80×10^{-3}	8.21×10^{-3}	1.53×10^{-2}	1.92×10^{-2}	

